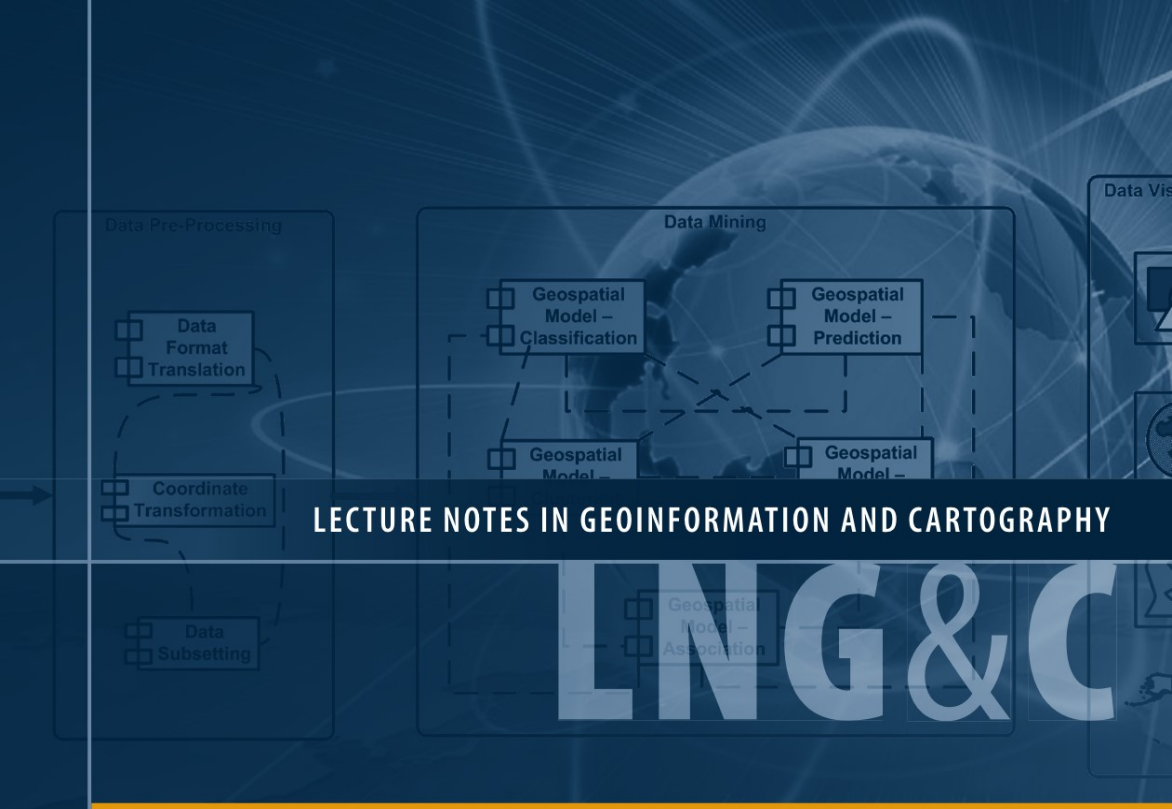

LECTURE NOTES IN GEOINFORMATION AND CARTOGRAPHY

LNG&C

Liping Di · H. K. Ramapriyan (Eds.)

Standard-Based Data and Information Systems for Earth Observation

Springer

Lecture Notes in Geoinformation and Cartography

For further volumes:
http://www.springer.com/series/7418

Liping Di · H.K. Ramapriyan
Editors

Standard-Based Data and Information Systems for Earth Observation

Springer

Editors
Dr. Liping Di
Center for Spatial Information
Science & Systems (CSISS)
Greenbelt MD 20770
USA
ldi@gmu.edu

Dr. H.K. Ramapriyan
NASA Goddard Space Flight Center
Greenbelt, MD 20771
USA
Rama.Ramapriyan@nasa.gov

ISSN 1863-2246 e-ISSN 1863-2351
ISBN 978-3-540-88263-3 e-ISBN 978-3-540-88264-0
DOI 10.1007/978-3-540-88264-0
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2009936962

Cover design: deblik, Berlin

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Contributors

Jeanne Behnke NASA Goddard Space Flight Center, Greenbelt, MD, USA, Jeanne.behnke@nasa.gov

Luis Bermudez Southeastern Universities Research Association (SURA), Newport News, VA 23606 USA, bermudez@sura.org

Joanne Bintz Southeastern Universities Research Association (SURA), Newport News, VA 23606 USA, bintz@sura.org

Philip Bogden Southeastern Universities Research Association (SURA), Newport News, VA 23606 USA, bogden@sura.org

B. Bonnlander Institute for Human and Machine Cognition, Pensacola, FL, USA, bbonnlander@ihmc.us

Helen Conover Information Technology and Systems Center, The University of Alabama in Huntsville, AL, USA, HConover@itsc.uah.edu

Constantino Cremidis CSC, Suitland, MD 20746, USA, constantino.cremidis@noaa.gov

D. Danks Carnegie Melon University, Pittsburgh, PA, USA, ddanks@cmu.edu

Liping Di Center for Spatial Information Science and Systems (CSISS), George Mason University, Greenbelt, MD 20770, USA, ldi@gmu.edu

Marilyn Drewry Information Technology and Systems Center, The University of Alabama in Huntsville, AL, USA, mdrewry@itsc.uah.edu

Ruth Duerr Cooperative Institute for Research in Environmental Science, National Snow and Ice Data Center, University of Colorado at Boulder, CO, USA, rduerr@nsidc.org

Yonsook Enloe SGT, Inc., Greenbelt, MD, USA, yonsook@mindspring.com

Mary Ann Esfandiari NASA Goddard Space Flight Center, Greenbelt, MD, USA, mary.a.esfandiari@nasa.gov

Sara Graves Information Technology and Systems Center, The University of Alabama in Huntsville, AL, USA, sgraves@itsc.uah.edu

Daniel Holloway OPeNDAP, Inc., Narragansett, RI, USA, d.holloway@opendap.org

D. Isaac BPS Consulting, Inc., Bethesda, MD, USA

John Jackson Nielsen Engineering & Research, Inc., Mountain View, CA, USA, jjackson@nearinc.com

Ken Keiser Information Technology and Systems Center, The University of Alabama in Huntsville, AL, USA, kkeiser@itsc.uah.edu

Dawn Lowe NASA Goddard Space Flight Center, Greenbelt, MD, USA, dawn.lowe@gsfc.nasa.gov

Manil Maskey Information Technology and Systems Center, The University of Alabama in Huntsville, AL, USA, mmaskey@itsc.uah.edu

Kenneth R. McDonald NOAA/NESDIS/OSD/TPIO, Silver Spring, MD 20910 (formerly with NASA/GSFC, Greenbelt, MD 20771), USA, Kenneth.Mcdonald@noaa.gov

Ross K. Meentemeyer Department of Geography and Earth Sciences, University of North Carolina, Charlotte, NC, USA, rkmeente@uncc.edu

Mark A. Parsons National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Science, University of Colorado at Boulder, CO, USA, parsonsm@nsidc.org

George Percivall Open Geospatial Consortium, Inc., Herndon, VA 20170-4819, USA, gpercivall@opengeospatial.org

Hampapuram K. Ramapriyan NASA Goddard Space Flight Center, Greenbelt, MD, USA, rama.ramapriyan@nasa.gov

Robert H. Rank NOAA/NESDIS/OSD/GSD/CLASS, Suitland, MD 20746, USA, Robert.Rank@noaa.gov

Laura C. Rodman Nielsen Engineering & Research, Inc., Mountain View, CA, USA, rodman@nearinc.com

Matt Smith Information Technology and Systems Center, The University of Alabama in Huntsville, AL, USA, msmith@itsc.uah.edu

Edwin Sofinowski SGT, Inc., Greenbelt, MD, USA, Edwin.J.Sofinowski@nasa.gov

Richard E. Ullman NASA/Goddard Space Flight Center, Greenbelt, MD, USA, Richard.E.Ullman@nasa.gov

Ron Weaver National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Science, University of Colorado at Boulder, CO, USA, weaverr@nsidc.org

Yaxing Wei Center for Spatial Information Science and Systems (CSISS), George Mason University, Greenbelt, MD 20770, USA, weiy@ornl.gov

Wenli Yang Center for Spatial Information Science and Systems (CSISS), George Mason University, Greenbelt, MD 20770; Fairfax, VA, USA, wyang1@gmu.edu

Le Yu Institute of Spatial Information Technique, Department of Earth Sciences, Zhejiang University, Hangzhou, 310027, P. R. China, naisoild@gmail.com

Genong Yu Center for Spatial Information Science and Systems (CSISS), George Mason University, Greenbelt, MD 20770, USA, gyu@gmu.edu

Peng Yue State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, 430079; Center for Spatial Information Science and Systems (CSISS), George Mason University, Greenbelt, MD 20770, USA, geopyue@gmail.com

Dengrong Zhang Department of Earth Sciences, Institute of Spatial Information Technique, Zhejiang University, Hangzhou, 310027, P. R. China, zju_rs@126.com

Peisheng Zhao Center for Spatial Information Science and Systems (CSISS), George Mason University, Greenbelt, MD 20770, USA, pzhao@gmu.edu

Chapter 1
Standards-Based Data and Information Systems for Earth Observations – An Introduction

Liping Di and H.K. Ramapriyan

In the past several years, we have witnessed an explosive growth in the information technology, especially the Internet and Web. Web services have become the mainstream practices in Web applications. Web service interoperability and service chaining have become a reality, thanks to the interoperability standards developed by the World Wide Web Consortium (W3C) and the Organization for the Advancement of Structured Information Standards (OASIS).

The geospatial community has followed the trend in the information technology world. The International Organization for Standardization (ISO) Technical Committee (TC) 211 has set a series of international standards on geographic information. The Open Geospatial Consortium (OGC), a non-profit, international, voluntary consensus standards organization, has developed a set of interoperability implementation specifications aimed at achieving the interoperability among geo-information systems and services. In particular, the OGC Web services specifications have been widely accepted by the geospatial and Earth observation communities. In the United States, the Federal Geographic Data Committee (FGDC) has also set a series of geospatial standards that are mandatory for the US federal agencies.

On the other hand, as a result of the recent advances in sensor and platform technologies, Earth observation sensors onboard different platforms have collected large volumes of geospatial data. The major form of geospatial data is remotely sensed imagery. A large number of countries have launched Earth observation satellites and numerous government agencies and private entities around the world have engaged in collecting, disseminating, and utilizing geospatial data. In the

L. Di (✉)
Center for Spatial Information Science and Systems (CSISS), George Mason University, Greenbelt, MD 20770, USA
e-mail: ldi@gmu.edu

This work was performed by the second author (Ramapriyan) as part of his official duties as an employee of the US government. It was supported by the NASA's Science Mission Directorate. The opinions expressed are those of the authors and do not necessarily reflect the official position of NASA.

L. Di, H.K. Ramapriyan (eds.), *Standard-Based Data and Information Systems for Earth Observation*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-540-88264-0_1, © Springer-Verlag Berlin Heidelberg 2010

United States, dozens of federal agencies have been working on the collection, management, archiving, processing, distribution, and application of geospatial data. For example, NASA's Earth Observing System (EOS) program alone has collected several petabytes of Earth observation data, and the amount is growing by a few terabytes a day.

With those in the background, the data and information systems for Earth observations have experienced some major changes in recent years. The advances in computer and information technologies have made the data more readily accessible through the Internet. Many data centers have setup data servers for Web-based on-line access to their data holdings. Instead of using proprietary standards and practices, newly developed and operated data systems have widely adopted the interoperability standards and specifications developed by ISO TC 211, OGC, and FGDC.

The large volume of geospatial data resources, the availability of on-line open data servers, and the existence of interoperability standards and technology form a common foundation for the sharing and interoperability of geospatial data, on which many value-added services and applications of national and international importance can be built. The question is how to use effectively such distributed, large, and valuable on-line geospatial data resources in applications. To be useful, these geospatial data must be further processed to extract application-specific geospatial information and knowledge. Traditionally, standalone geographic information systems (GIS) and image exploitation systems have been used for information extraction and knowledge discovery from the geospatial data. Such systems require the trained experts to operate. With the limited availability of trained professionals, use of geospatial data in applications has been very expensive. The geospatial web service and geospatial semantic web technology provides the promise for greatly facilitating and even automating some of the processes of converting data to user specific knowledge. The geospatial web services are based on a service-oriented architecture (SOA), in which, individual data analysis functions become standard-compliant and chainable web services can be distributed over the web. Those services can be chained together dynamically to solve complex geospatial problems. The geospatial semantic web technology, with the support of geospatial ontologies, provides the means to automatically discover and chain the relevant services and data. With such technologies, geospatial knowledge building systems can be built to provide geospatial knowledge services to wide user communities.

This book summarizes the recent advances in the standard-based data and information systems for Earth observation. The topics covered by the book include new or updated standards, development of standards-based data systems, data access and discovery services, new data capabilities and sources available through standards-based data systems, data systems architecture, lessons learned from development, deployment, and operation of standards-based data systems, and other related subjects. In addition, this book also addresses the new technologies for SOA-based geospatial knowledge systems, including knowledge discovery algorithms, distributed image information mining, architectures and standards, knowledge system prototypes, geospatial knowledge representation, and geospatial semantic Web.

This book is organized as follows: Chapters 2–6 address the latest development of geospatial interoperability standards and the implementation and evolution of data systems with the latest standards and technologies. Chapters 7–8 discuss the advances in the standard-based preservation and archival of geospatial data and metadata. Chapters 9–12 cover the topics of knowledge discovery, data mining, and semantic Web. Chapter 13 discusses the strategy and approach for accelerating technology adoption. The following paragraphs discuss the content of each chapter briefly.

Spatial Information Grid (SIG) is a fundamental infrastructure for sharing and interoperation of distributed geospatial data. Many national and international organizations are constructing SIGs. Chapter 2, authored by Zhang et al., discusses the experience and lessons-learned from the implementation of a data node in China's SIG by using the OGC standards compliant NASA HDF-EOS Web GIS Software Suite (NWGISS).

Data access and analysis tools that are developed within specific disciplines, and the protocols that they are built upon provide valuable services to their respective users but can actually be a barrier to the integration of data from a broad set of data sources. An example of this is the difficulty encountered in integration of data supported by OPeNDAP that is widely used in the ocean and atmospheric sciences, and data provided through the interface specifications of the Open Geospatial Consortium (OGC) that typically serves the land science community. Chapter 3, by McDonald et al., describes a project that has developed a gateway to bridge these two data system infrastructures, in response to a specific need expressed by Coordinated Enhanced Observing Period (CEOP), an international science program.

Standards are the key for achieving the interoperability and sharing of geospatial resources. OGC is a major player in setting implementation standards for geospatial interoperability. Chapter 4, authored by George Percivall, presents OGC's vision on geospatial interoperability, reviews the OGC organization and standards, and discusses in detail the progress on OGC Web service interoperability.

Space agencies around the world are operating a number of large data systems to support their satellite-based Earth observations. Many of such systems were built several years ago with large investments. Evolution of such systems with advances in the geospatial information technologies is a major concern. One example of such systems is NASA's Earth Observing System (EOS) Data and Information System (EOSDIS), which has been serving a broad user community since 1994. Most of NASA's Earth science data are currently being archived, managed and distributed by EOSDIS. As of the end of 2007, the archives of EOSDIS held over 3.7 petabytes of data from over 90 instruments and over 2000 distinct science products. The distribution of data to end users in 2007 amounted to approximately 4 TB a day. The community receiving data from EOSDIS is on the order of 200,000 distinct users from a diverse set of organizations and scientific disciplines. While EOSDIS is effectively managing a large amount of data and successfully serving a broad user community, it is a system whose design and development originated more than 15 years ago during which many advances have occurred in information technology. Chapter 5, by Ramapriyan et al., discusses NASA's approach to and lessons learned from the EOSDIS evolution through technology infusion to increase end-to-end data

system efficiency and autonomy while decreasing operations costs, increase data interoperability and usability by the science research, application, and modeling communities, improve data access and processing, and ensure safe stewardship.

Chapter 6, by Conover et al., discusses a standards-based distributed information system for coastal data management, which has been developed for the Southeastern Universities Research Association (SURA) Coastal Ocean Observing and Prediction (SCOOP) program. SCOOP is a distributed program, incorporating heterogeneous data, software and hardware; thus the use of standards to enable interoperability is key to SCOOP's success. Standards activities range from internal coordination among SCOOP partners to participation in national standards efforts. A suite of advanced technologies have been developed to provide core data and information management services for scientific data, including the SCOOP Catalog and a suite of standards-based web services for data discovery, access and visualization.

Metadata are the data about data. Metadata are very important for scientific data since metadata provide the quality, usage, lineage, and other information of the scientific data. Therefore, it is important to preserve the metadata. Chapter 7, A New Approach to Preservation Metadata for Scientific Data-A Real World Example, by Duerr, et al., describes the US National Snow and Ice Data Center (NSIDC)'s efforts and lessons-leaned on the development of a prototype operations and preservation metadata tool based on the OAIS Reference Model and compatible with the PREservation Metadata Implementation Strategies (PREMIS) Data Dictionary in order to consolidate the operations and preservation metadata collected for many of NSIDC's datasets.

Chapter 8, Archive Standards: How Their Adoption Benefits Archive Systems, by Rank et al., discusses how the adoption of standards in general and a submission process developed using the recommendations for Space Data System Standards from the Consultative Committee for Space Data Systems (CCSDS) reference model for and Open Archival Information System (OAIS) has supported the development and operations of the Comprehensive Large Array-data Stewardship System (CLASS) of the US National Oceanic and Atmospheric Administration (NOAA).

In order for geospatial data to be useful, they need to be further processed to extract information and knowledge for supporting applications and decision making. The advances of the last two decades in remote sensing instruments, computational, storage and communications hardware, and launches of a series of Earth observing satellites by US and international agencies, have created a data rich environment for scientific research and applications. Chapters 9 and 10 describe some of the latest developments in knowledge discovery from geospatial data available through this environment. Chapter 9, An Association Rule Discovery System for Geographic Data, by Rodman et al., presents an association rule discovery system, called Aspect, for discovering knowledge from geospatial data. The system works with standard geographic data formats and extends the association rule formulation to handle spatial relationships. Chapter 10, An Intelligent Archive Testbed Incorporating Data Mining, by Ramapriyan et al., discusses a large-scale testbed that looks upon the distributed provider environment with capabilities to convert data to information

and to knowledge through data mining as an Intelligent Archive in the Context of a Knowledge Building system (IA/KBS). There have been several research investigations into intelligent data understanding including data mining and knowledge discovery. However, these investigations typically perform proofs of concept on a relatively small scale. Before their contributions can be implemented on a large scale commensurate with today's Earth science data archives, it is necessary to test them in a pseudo-operational environment. The testbed serves this purpose and the chapter provides a discussion of some of the observations and lessons learned from its implementation.

The semantic Web is one of today's hot research areas in the information technology. In the semantic Web, the semantics of information and services on the Web are defined in a standard way so that the consumers of the information and services (either people or machines) can understand and meaningfully use the Web content. Geospatial semantic Web is the application of semantic Web technology in the geospatial domain. Chapters 11 and 12 provide examples of some of the latest geospatial semantic Web studies.

Chapter 11, Semantic Augmentations for Geospatial Catalogue Service, by Yue et al., discusses a research on the geospatial semantic catalogue. Catalogue service plays an important role in helping requestors to find the suitable geospatial data and services over the Web. The OGC has developed and recommended an ebRIM profile of Catalogue Services for the Web (CSW) for implementing a catalogue service. Metadata for data and services registered in CSW are usually described by following the existing geographic metadata standards. The search functionality is limited to the direct match of keywords from metadata without fully utilizing the semantic information implicitly embedded in the metadata, such as hierarchical relationships among metadata entities. Web Ontology Language (OWL) provides a mechanism to enable the use of semantics. OWL-S uses OWL to describe the semantics for Web service. Chapter 11 explores the semantic representation of geospatial data and services to enable the semantic search in CSW based on the semantic relationship defined in OWL/OWL-S. Such semantics are organized in CSW through extending ebRIM elements. The chapter also illustrates how such semantically augmented CSW can facilitate service chaining and assist in dynamic discovery and/or derivation of geospatial information.

Chapter 12, Geospatial Knowledge Discovery Using Semantic Web Services, by Zhao and Di, explores the application of geospatial semantic Web technology in the knowledge discovery. Large amount of Earth and space science data has been and continue to be collected from various sources. Effective and efficient knowledge discovery from these distributed multi-disciplinary and multi-scale data is becoming a big challenge. It requires the relevant data and processing steps' being discovered, accessed and integrated as much as possible. The Semantic Web provides a common interoperable framework in which information is given well-defined meaning such that the data and operations can be used for more effective discovery and integration across various applications. This paper introduces a new approach to distributed data mining for geospatial knowledge discovery based on semantic Web services and their automatic and semi-automatic chaining. In this approach, domain concepts are

well defined by geospatial ontology as the basic knowledge, data and data mining processes then are well described by these concepts and served by OGC Web services and semantic Web services. So the whole process of geospatial knowledge discovery can be represented as a service chain in predefined patterns of domain concepts. This approach provides an infrastructure that enables individual data and data mining software not only discoverable and accessible, but also interoperable in order to assemble them automatically or semi-automatically to implement more complicated geospatial knowledge discovery.

We conclude the book with the Chapter 13, Accelerating Technology Adoption through Community Endorsement, by Ullman and Enloe. As we discussed at the beginning of this introduction, the information technology has been advancing rapidly in the recent years. However, the adoption of the new technology by the Earth science community has not been as rapid as expected. This chapter discusses a process employed by the Standards Process Group (SPG), one of four NASA Earth Science Data Systems Working Groups, to accelerate the technology adoption. The center of the process is the Request for Comments (RFC) model successfully used by the Internet standard-setting organizations. The purpose of the RFC is to notify the wider community of specific detailed ideas that potentially affect interoperation of geospatial data and services though the Internet. Through the RFC mechanism, ideas from the stakeholders are shared, adoption of new technologies spurred, and collaboration in the development of geospatial standards fostered.

Chapter 2
Use of NWGISS to Implement a Data Node in China's Spatial Information Grid

Dengrong Zhang, Le Yu, and Liping Di

2.1 Introduction

Geospatial data are those that can be associated with location information on the Earth. Because of their importance, both public and private organizations have collected considerable geospatial data (King 1999, King et al. 2003, McDonald and Di 2003). Such data is the dominant form in volume. It has been widely used in many fields of applications. China has accumulated large-scale, heterogeneous spatial resources, among them, continuing to establish a fundamental spatial database, spatial data processing and application software, spatial facilities, and instruments (Guo et al. 2004). Consistent access and sharing of spatial information are generally considered to be challenging problems due to the volume and complexity of processing heterogeneous and distributed data. Technology for extensive GIS application is needed to implement effective access and sharing of the large amount of isomerous and distributed spatial data.

Many approaches have been applied to implementing sharing and integration of spatial information, Grid technology, OpenGIS and Web services are the three most important (Tang and Jing 2004). The Grid technique, first proposed by Ian Foster et al. (2001), is a flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources. It has been a rising research field in recent years (Shao and Li 2005). There have been some studies on integrating grid technology with spatial information applications. The Committee on Earth Observation Satellites (CEOS) started research in 2001 (Tang and Jing 2004) on a prototype system to share satellite data and spatial information in global areas. In China, the National University of Defense Technology (NUDT) first proposed SIG as a system, to integrate grid technology with spatial information applications. It conducted original and fundamental research on SIG architecture. Beginning in

D. Zhang (✉)
Department of Earth Sciences, Institute of Spatial Information Technique, Zhejiang University, Hangzhou, 310027, P. R. China
e-mail: zju_rs@126.com

L. Di, H.K. Ramapriyan (eds.), *Standard-Based Data and Information Systems for Earth Observation*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-540-88264-0_2,

2002, the National High Technology Research and Development 863 Program of China supported a prototype system.

This chapter concentrates on resource management, because it is fundamental to other SIG services. The data node is designed to be the unit for resource management. Given the requirement of a SIG data node, and using the architecture of the NASA HDF-EOS Web GIS Software Suite (NWGISS) (http://nwgiss.laits.gmu.edu/introduction.htm, Yang and Di), which is a Web-based data distribution system compliant with multiple Open Geospatial Consortium (OGC) standards and the OGC Web Services (OWS) frame, a SIG data service node framework was designed. This paper describes it. Test geospatial data nodes based on the SIG data node prototype have been established for evolution at Zhejiang University and George Mason University.

SIG and OGC standard data interoperability protocols are introduced in Sect. 4.2. Section 4.3 focuses on SIG resources and the SIG data node. Section 4.4 describes the construction of an NWGISS-based SIG data node. An experimental distributed evolution framework and a demonstration of an application are given in Sect. 4.5.

2.2 Related Work

2.2.1 Introduction to SIG

As a novel Web-based infrastructure and technology system of spatial information, SIG integrates and extends information grid technology, spatial information systems and Web services. To implement sharing and integration of spatial information, SIG takes services as its technical core and establishes a unified and intelligent platform to acquire, store, organize, distribute, analyze, aggregate, and apply spatial information.

SIG provides solution to problems in and meets the needs of spatial information application. These problems and needs are focused mainly on (Ren et al. 2004) integrative organization of spatial information resources, sharing large amounts of spatial information resources, high performance collaboration in analyzing and processing spatial information, and integration of geographically distributed spatial information services. Luo et al. (2004) summarized seven major functions SIG should provide:

(1) Ability to process massive amounts of spatial data. Storing, accessing and managing spatial data in amounts from terabytes to petabytes; efficiently analyzing and processing spatial data to produce models, information, and knowledge; and providing 3D and multimedia visualization services.
(2) High performance computing with and processing of spatial information. Solving spatial problems with high precision, high quality, and on a large scale; and processing spatial information in real time or on schedule, with high-speed and high efficiency.

(3) Sharing of spatial resources. Sharing distributed heterogeneous spatial information resources and realizing interlink and interoperation at the application level, so as to make the best use of such spatial information resources as computing resources, storage devices, spatial data (integrating from GIS, RS and GPS), spatial applications and services, GIS platforms (such as ESRI ArcInfo, MapInfo, . . .).
(4) Integration of legacy GIS systems. A SIG can be used not only to construct new advanced spatial application systems, but also to integrate legacy GIS systems, to keep extensibility and inheritance and guarantee the users' investment.
(5) Collaboration. Large-scale spatial information applications and services always involve different departments in different geographic locations, so remote and uniform services are needed.
(6) Support to integration of heterogeneous systems. Large-scale spatial information systems are always synthesized applications, so SIG should provide interoperation and consistency through adopting open and applied technology standards.
(7) Adaptability to dynamic changes. Business requirements, application patterns, management strategies, and IT products for any department are always changing, so SIG should be self-adaptive.

Tang and Jing (2004) first proposed the architecture of SIG. The main components of SIG are systems that acquire spatial information, storing systems, processing systems, application systems, multi-layer users, and computing resources (e.g. PCs, servers). These components are linked and integrated by SIG services (see Fig. 2.1).

There are many reasons why one might wish to have SIG. First, the amount of spatial data is increasing amazingly, so that real time or near real time processing needed by applications confronts difficulties in one single computer. Second,

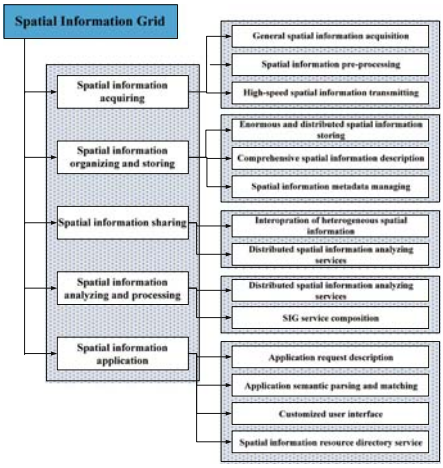

Fig. 2.1 The technical architecture of SIG (Tang and Jing 2004)

data, algorithms, and/or computing resources are physically distributed. Third, the resources may be "owned" by different organizations. Fourth, the use frequency of some resources is rather low. A SIG contains at least (Luo et al. 2006):

(1) A Remote Sensing Information Analysis and Service Grid Node
(2) A data service node: the traditional data base for a Web service
(3) A management center: resource register, finding data and services, services trading, and management;
(4) A portal: an entry for SIG users.

2.2.2 Standards for Geospatial Data Operation

Many international and industry standards have been established to implement Web-based interoperable geospatial information data access and services for GIS research. ISO TC211 (http://www.isotc211.org/Outreach/Overview/Overview.htm) standards and Open GIS Consortium (OGC) specifications (http://www.opengeospatial.org/standards/as) are the most attractive. ISO TC211 is a technical committee with responsibility for establishing the international standards for geographic information. The OGC is a not-for-profit international membership-based organization founded in 1994 to address the lack of interoperability among systems that process geo-referenced data. OGC advances geospatial interoperability technology by developing interoperable interface specifications. Those specifications, tested through interoperability initiatives, are widely accepted by software vendors, the GIS community, and federal agencies in the US. They are also adopted by many different countries and international organizations. The Memorandum of understanding signed by ISO TC 211 and OGC states that OGC will submit its specifications to ISO TC 211 for approval as international standards. TC 211 usually accepts these documents as Committee Drafts.

Among all OGC specifications the most significant kernel specifications for establishing spatial information grid data nodes are Web Coverage Service (WCS) (http://www.opengeospatial.org/standards/wcs),Web Feature Service (WFS) (http://www.opengeospatial.org/standards/wfs), Web Map Service (WMS) (http://www.opengeospatial.org/standards/wms), Catalog Service/Web Profile (CS/W) (http://www.opengeospatial.org/standards/cat), and Geography Markup Language (GML) (http://www.opengeospatial.org/standards/gml). WCS and WMS provides an interoperable way for accessing geospatial overlay data. WFS provides an interoperable approach for accessing geospatial feature data. CS/W defines an interface for Web geospatial data query access. GML provides a tool for describing geospatial data. These kernel specifications form the interoperable bases of geospatial data. WCS, WFS and WMS are three specifications that most related to data sharing and interoperation.

2.2.2.1 OGC WCS

The OGC WCS specification defines the interface between Web-based clients and servers for interoperable access to on-line multi-dimensional, multi-temporal geospatial data (_http://www.opengeospatial.org/standards/wcs). According to definitions by OGC, coverage data include all remote sensing images as well as gridded data such as DEM and land use classification. Three operations are defined in WCS:

GetCapabilities: Client retrieves the XML-encoded capabilities document from a server. The document contains information about the data it serves, as well as about the server capabilities.

GetCoverage: Client requests the server to send data based on client's requirements.

DescribeCoverage (optional): Client retrieves the metadata for a specific coverage.

2.2.2.2 OGC WFS

The OGC WFS specification defines the interfaces between Web-based clients and servers for accessing feature-based geospatial data (http://www.opengeospatial.org/standards/wfs). Examples of geospatial feature data are transportation road networks, coastlines, political boundaries, and utility lines. The WCS and WFS together provide standardized, on-line access to all geospatial data. They form the foundation for Web-based interoperable access of geospatial data. The WFS specification defines three mandatory operations for accessing and manipulation of feature data:

GetCapabilities: A Web feature server must be able to describe its capabilities. Specifically, it must indicate which feature types it can serve and what operations on each feature type are supported.

Get Feature: A Web feature server must be able to service a request to retrieve feature instances. In addition, the client should be able to specify which feature properties to fetch and should be able to constrain the query spatially and non-spatially.

DescribeFeatureType: A Web feature server must be able, upon request, to describe the structure of any feature type it can serve.

2.2.2.3 OGC WMS

The OGC WMS specification defines Web interfaces for dynamically assembling maps over the Internet from multiple sources within a heterogeneous distributed computing environment (http://www.opengeospatial.org/standards/wms). Maps are the visualization of data. A WMS server normally converts the data in its archive to a visualized form (map) based on the requirements of the client. In many cases, a WMS server may talk to a WCS or WFS server to obtain the needed data for making

maps requested by a client. In this sense, a WMS server can be considered as a data visualization service for either WFS or WCS servers. The WMS specification defines three operations:

GetCapabilities (required): Obtain service-level metadata, which is a machine-readable and human-readable description of the WMS's information content and acceptable request parameters.
GetMap (required): Obtain a map image whose geospatial and dimensional parameters are well defined.
GetfeatureInfo (optional): Ask for information about particular features shown on a map.

2.3 SIG Resources and Data Node

2.3.1 SIG Resources

The SIG framework can be divided into four layers (Tang et al. 2004): the resource layer, share layer, assembly layer and application layer. The resource layer is composed of the distributed geospatial file server, geospatial database server, remotely sensed imagery server, and sensor simulation node. All distributed resources are packaged and connected to the SIG system by the SIG Resource Package. The share layer is composed of the SIG Resource Management Service and SIG Resource information Service. It organizes, collects, discovers, and selects global spatial information resources. The assembly layer is composed of the SIG Spatial Information Resource Engine and SIG Information Index Engine. The former assembles and combines business logic, the latter searches quickly. The application layer realizes a SIG Portal for spatial users.

To make use of a spatial Web service, the user needs an interpretable and standard description and the means by which the service is accessed. An important goal in managing spatial resources information is to establish a framework within which these descriptions are made and shared. Besides technical support, the framework provides a unified starting point, which is the resources information registry for resources information description, publication, discovery, and employment (Guo et al. 2004).

The resource layer is the fundamental resource environment of SIG. It can be denoted by R (W, S, D), where W= {Image Service, Map Service, Feature Service. . .} represents the services that can be provided, S= {Data Description and Architecture Protocol, Data Access Protocol, Service Interoperation Protocol} represents the protocols that should be followed and D= {Spatial Data Files, Spatial Database} represents the data that can be provided. Of the three elements, W and D are resources for computing and data, while S is the access rule linking W and D. Figure 2.2 shows the concept model of the resource layer.

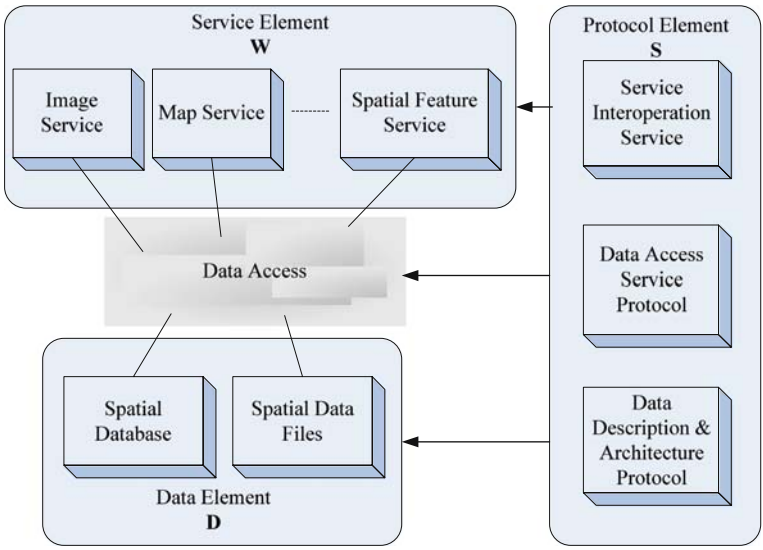

Fig. 2.2 SIG resource layer concept model

2.3.2 SIG Data Node

SIG's resource infrastructure node, which is composed of software, data, hardware, and protocols, can provide spatial data services to Grid. Furthermore, the data node follows OGC specifications in interface and interoperation.

The function of the SIG data node is to organize spatial data from distributed spatial data file servers, spatial database servers, and remotely sensed imagery servers into specific structures for easier managing, access, and use. So there are two requirements for this node:

(1) It should provide coverage services, map services, feature services, and catalog services.
(2) All of these services should comply with OGC specifications.

2.4 SIG Data Node Based on NWGISS

2.4.1 Structure of NWGISS

NWGISS was developed by the Laboratory for Advanced Information Technology and Standards (LAITS) at George Mason University (GMU) to manage the large volume of HDF-EOS format remote sensing data generated by NASA's Earth

Science Enterprise (ESE). LAITS' NWGISS design, a Grid based three-layer structure (Di et al. 2003), significantly improves the accessibility, interoperability of HDF-EOS data. It works with all HDF-EOS files. LAITS's NWGISS is also a Web-based data distribution system, compliant with multiple OGC standards. NWGISS consists of the following components: a Web map server (WMS), a Web coverage server (WCS), a Catalog Service/Web Profile (CS/W) server, a multi-protocol geo-information client (MPGC), and a toolbox (http://laits.gmu.edu/DownloadInterface.html). All NWGISS components can work either independently or collaboratively. WCS and WMS were designed for the distribution of remote sensing data. The CS/W server provides general OGC catalog services. MPGC is a comprehensive OGC client. Currently, OGC WRS, WMS, WFS, and WCS have been implemented in the client. The interaction between MPGC and OGC-compliant Web servers provides interoperable, personalized, on-demand data access and services for geospatial data. The NWGISS architecture can be seen in Fig. 2.3. Functions of five components are listed below (Di et al. 2002):

1) Map Server: The map server enables GIS clients to access HDF-EOS data as maps. Currently, the NWGISS map server complies with OGC WMS version 1.1.0. The OGC specification defines three interfaces: GetCapabilities, GetMap,

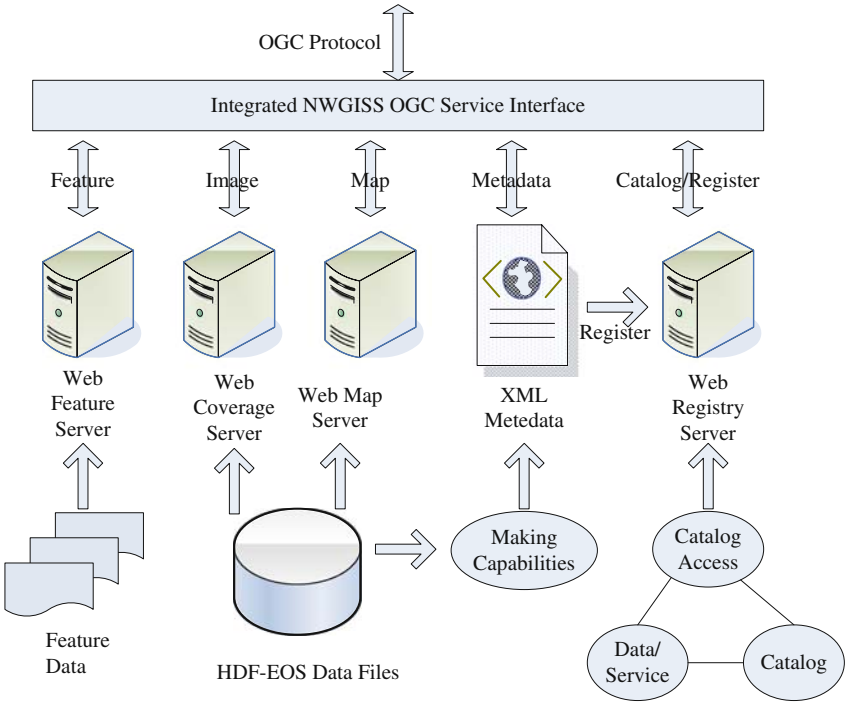

Fig. 2.3 NWGISS interoperation data server layer architechture (Tang and Jing 2004)

and GetFeatureInfo. All three interfaces have been implemented and all three HDF-EOS data models (Grid, Point, and Swath) are supported.

2) Coverage Server: The OGC Web Coverage Service (WCS) specification is designed to enable GIS clients to access multi-dimensional, multi-temporal geospatial data. WCS defines three interface protocols: getCapabilities, getCoverage, and describeCoverageType. The NWGISS coverage server has implemented both versions 0.5 and 0.6 of the draft WCS specification. NWGISS can return coverage in three formats: HDF-EOS (http://hdfeos.gsfc.nasa.gov/hdfeos/he4.cfm), GeoTIFF (http://remotesensing.org/geotiff/spec/geotiffhome.html), and NITFF.
3) Catalog Server: Both the WCS and WMS clients have the GetCapabilities protocol for finding geographic data/maps and services available from servers. This protocol works nicely when a server has a small data archive. If the server has a large quantity of data, the capabilities description, which basically is a data catalog, becomes very large. The catalog server allows GIS clients to search and find available geographic data and services in a NWGISS site following the OGC catalog interoperability specification (CIS). Both state-full and state-less OGC CIS have been implemented in the NWGISS catalog server, which reuses part of the Data and Information Access Link (DIAL) catalog server (Di et al. 1999).
4) Web Coverage Client: The NWGISS coverage client is a comprehensive OGC WCS client. It is able to interactively communicate with all OGC-compliant coverage servers for accessing multi-dimensional geospatial data and handling all three coverage-encoding formats, not only with NWGISS. Besides performing basic WCS client-server communication, coverage access, visualization, and user interaction, the client will also provide georectification, reprojection, and reformatting functions. The user's data requirement and the information about the data in the servers will automatically trigger execution of those functions, when required. The interaction between the NWGISS Web coverage client and OGC compliant Web coverage servers will provide interoperable, personalized, on-demand data access and services.
5) Toolbox: It contains tools for automated data ingestion and catalog creation. Currently, two types of tool are provided: format conversion tools and XML capabilities creation tools. A third type of tool the catalog creation tools, will be provided in the future.

2.4.2 Structure of SIG Data Node

To satisfy the requirements of a SIG data node, a framework of a SIG data service node was designed using NWGISS architecture and the OWS frame in Windows 2000 server and Linux operating system environments. A SIG data node architecture is shown in Fig. 2.4.

NWGISS is the core of the SIG data node. Format transfer is used to transfer other formats (such as GeoTiff) to HDF-EOS. Capabilities in WCS or WMS are generated from HDF-EOS data. These metadata documents will be registered in the

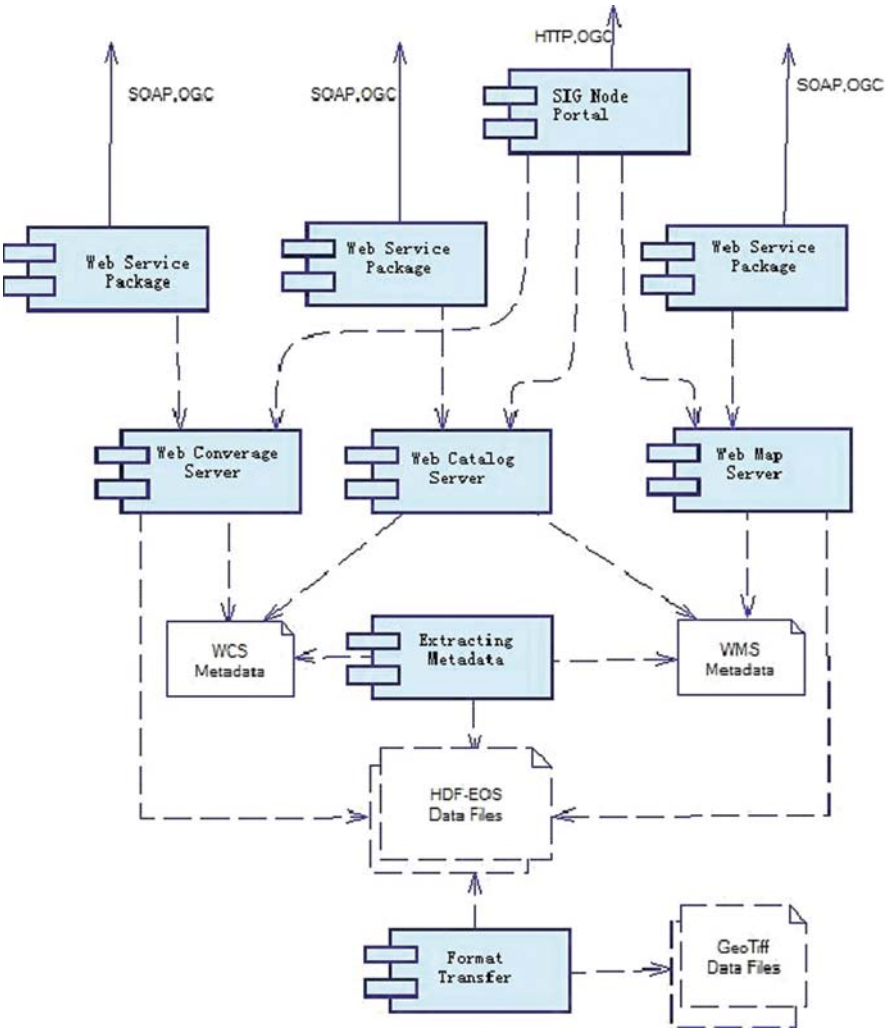

Fig. 2.4 SIG data node architecture

catalog server, available if required by the application layer or resource management layer.

The SIG node portal is a Website gateway for the service node. It is linked to all services by following OGC specifications. Clients from the application layer communicate with Portal by HTTP protocol. Base services, such as WCS, WMS, and CS/W, can interoperate directly with the resource or application layer. In order to satisfy the requirement of applications based on Web services, AXIS (The Apache Software Foundation) is used to convert (package) all services to Web service.

2.4.3 Geospatial Information Index

In order to efficiently support searching of geospatial data by different platforms, a quad-tree index structure was employed with OGC WCS specifications on both the Java and .net environments. Implementation of WCS in .net is based on .Net framework 2.0. It uses the C# 2.0 language to produce a WCS Web Service following the OGC standard. As shown in Fig. 2.5, the .net C# index class is composed of RequestParser.cs, DataIndex.cs, and Image.cs. RequestParser.cs is used to analyze parse Layer, BBOX, Width/Length, Format and to determine the spatial data index and the return data format. The main function of DataIndex.cs is to index the entire document. It first uses the Layer name that the user requested to determine the folder name; under this folder, it then uses the ratio of the Width/Length and the BBOX to determine the corresponding folder of images with the same spatial resolution. It then uses the BBOX to determine the document size and requested scope under this folder. Image.cs is used to splice the images that arrive to the index, return the mosaic image in the format requested by the user. The index process is invoked at GetCoverage time. Part of the analysis is completed in GetCoverage.cs, including request service name, edition.

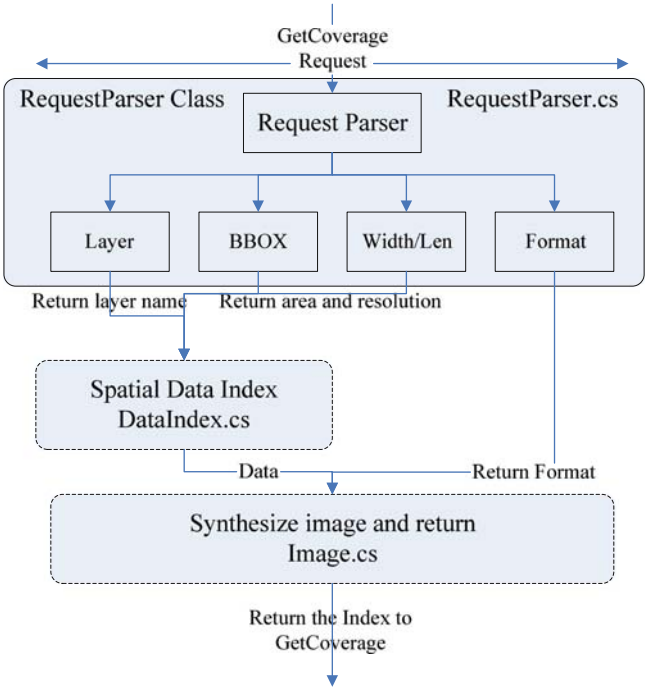

Fig. 2.5 Workflow for data index

On the other hand, implementation of WCS in Java is based on JRE 1.50. The Java development environment has already been installed, with the Web server being Tomcat and Web Service deployment by Apache AXIS. Complete spatial data index frameworks on Java and .net platforms are shown in Fig. 2.6.

This framework demonstrates the WCS flows using JAVA and .NET. The customer first uses GetCapabilities and DescribeCoverage requests to obtain the grid data description. Acting according to this description, the user then invokes the basic spatial data index mechanism, using GetCoverage to obtain the index to the grid document and return the grid image that the user requested.

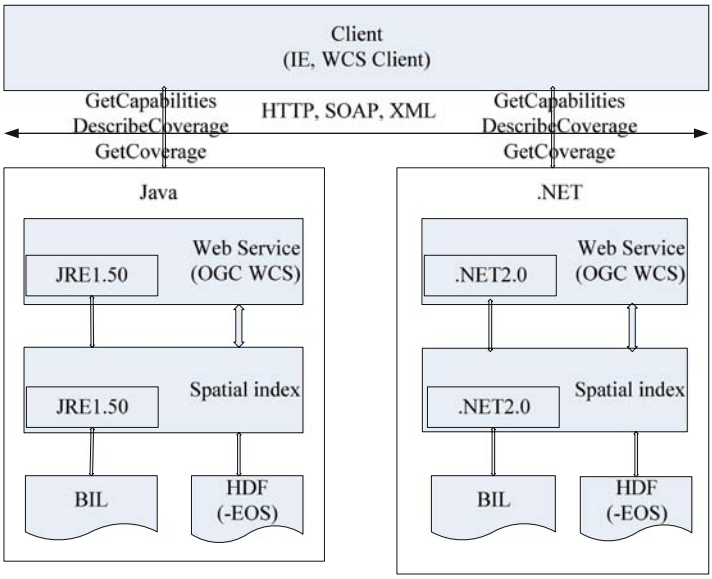

Fig. 2.6 System framework

2.5 Clients Implementation

2.5.1 Experiment Platform

The experiment platform (see Fig. 2.7) for SIG was built on Aparche/Tomcat in the Linux and Windows systems at the Institute of Spatial Information Technique (ISIT), Zhejiang University (ZJU) and GMU LAITS. The experiment concentrated on the resources layer.

ZJU ISIT manages the data node and provides access to the clients. It has responsibility for providing recent MODIS data and stored remote sensing images to SIG. This node server takes GMU LAITS's NWGISS as core software. The OGC WEB Coverage Service (WCS) 1.0 standard was employed as the data service interface. Any WCS1.0-compliant client may get data from the node. All the

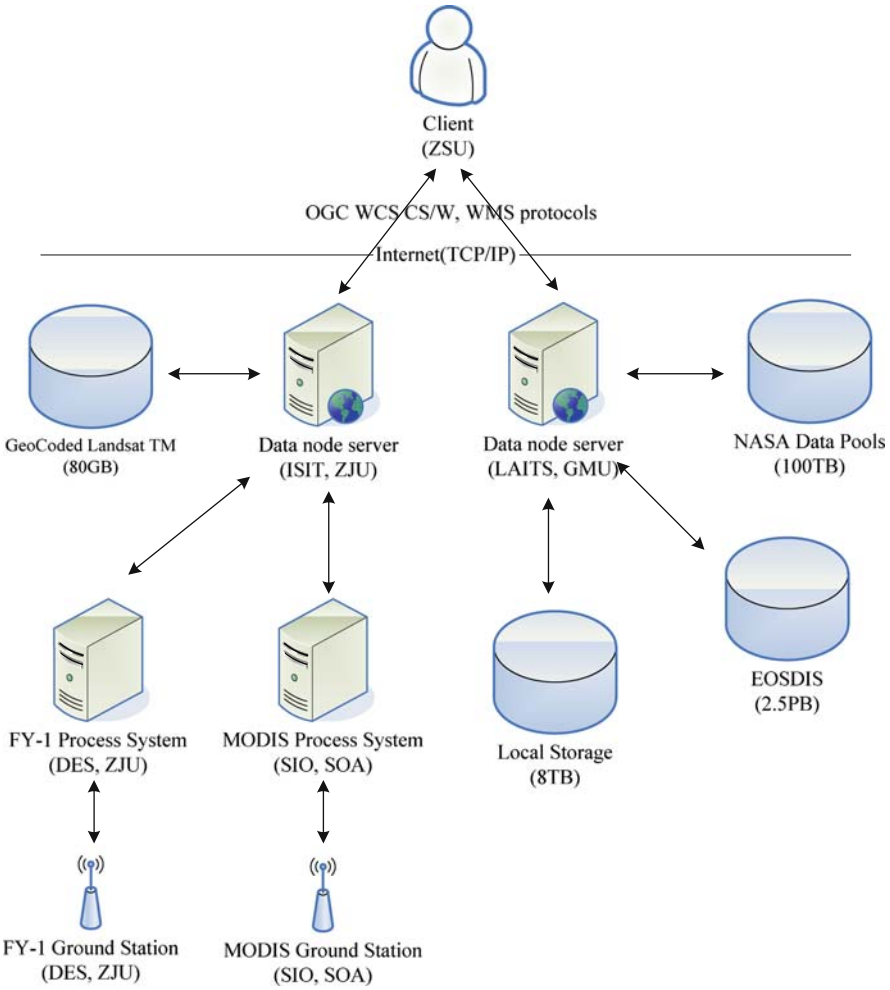

Fig. 2.7 Architecture of experimental system

location, time, projection, resolution, and format can be defined by the client. Three returned formats for coverage, HDF-EOS4, GeoTiff, and NITFF are supported by NWGISS.. WCS can also provide binary format data. The International Standard ISO 19115:2003 (http://www.isotc211.org/Outreach/Overview/Factsheet_19115.pdf) was employed to describe the dataset and ISO 19119:2005 (http://www.isotc211.org/outreach/overview/Factsheet_19119.pdf) was employed to register the data access service in the data node. Some metadata was also provided by the WCS XML Capability document. All the metadata can be obtained though WCS GetCapabilities and DescribeCoverage. The data query service is based on OGC Catalog Service (CS)/WEB (OGC CS/W) specifications. The node can provide

almost real-time MODIS Data, which the Second Institute of Oceanography (SIO), State Oceanic Administration (SOA), China receives and processes. MODIS data are available in SIG WCS 1~2 h after the satellite passes over the territory. This SIG data node also connects with the remote sensing data processing system for the FY-1 meteorological satellite, managed by the Department of Earth Sciences (DES), ZJU. A JSP service gate portal that includes all the services provided by the SIG data node was designed in order to provide convenient access to the data node service. The GMU LAITS standards node can directly connect with the NASA EOSDIS system, and it can provide large quantities of NASA EOS earth observation data through the standard interface.

2.5.2 Client Access Mode

All the interfaces of the SIG data node are compliant with OGC specifications. Three access modes were adopted for different applications: Web service mode, Web application mode and desktop application mode.

Web service mode: The Web Service package tool AXIS, is used to package OGC services at each data node. After packaging, three methods, including GetCapabilities, GetCoverage and DescribeCoverage, are designed to interoperate with the client using the SOAP protocol. Parameter formats in these methods are consistent with OGC WCS specification.

Web application mode: An OGC-based request string construct in a specific format is used to get geospatial information through Internet Explorer. "HTTP://3.40.56/cgi-bin/wcs?version=1.0.0&service=wcs&request= GetCapabilities" is the string to get WCS metadata at a data node whose IP is 10.13.40.56.

Desktop application mode: In this mode, OGC specifications are protocols for requests for services and determine application-programming interfaces (APIs). Therefore, in software systems with different purposes, OGC-based codes can be embedded in HTTP requests to get data properly from the node.

Since the WCS service complies with OGC specifications, any methods or client tools compliant with OGC standards can directly access the WCS services. Both Web applications and desktop applications invoke the SIG Web service; they act as clients in Web service mode.

2.5.3 Desktop Application Mode

GMU MPGC is a windows-based Java client, compliant with OGC specifications; it can access geospatial data via desktop application mode (see Fig. 2.8). The client is deployed at Zhejiang Shuren University (ZSU), China, and acts as an instance of the SIG application layer. It uses a friendly user interface to receive user request parameters through an http request that uses internal packaging for SIG services to access the data. The SIG data node at ZJU responds to MPGS, returning information as requested (see Fig. 2.9).

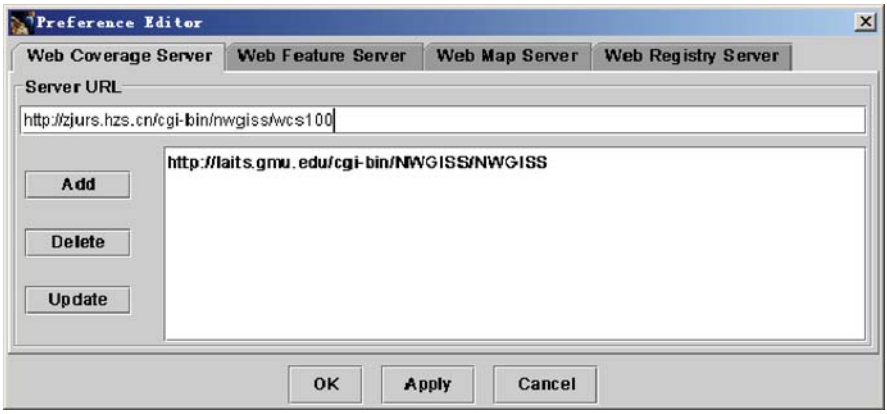

Fig. 2.8 MPGS request interface

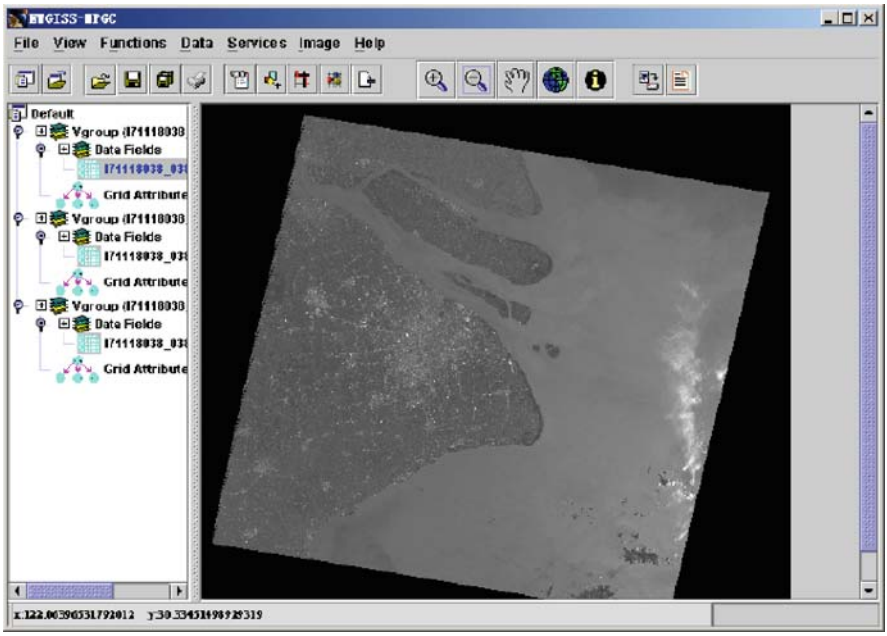

Fig. 2.9 Result from MPGS

2.5.4 Web Application Mode

Another case is a service chain application for water information extraction from Landsat TM. Figure 2.10 shows the framework.

This prototypes the integration of workflow technology, Web services, and the OGC Web Process Service specification (Version 1.0.0. http://www.opengeospatial.

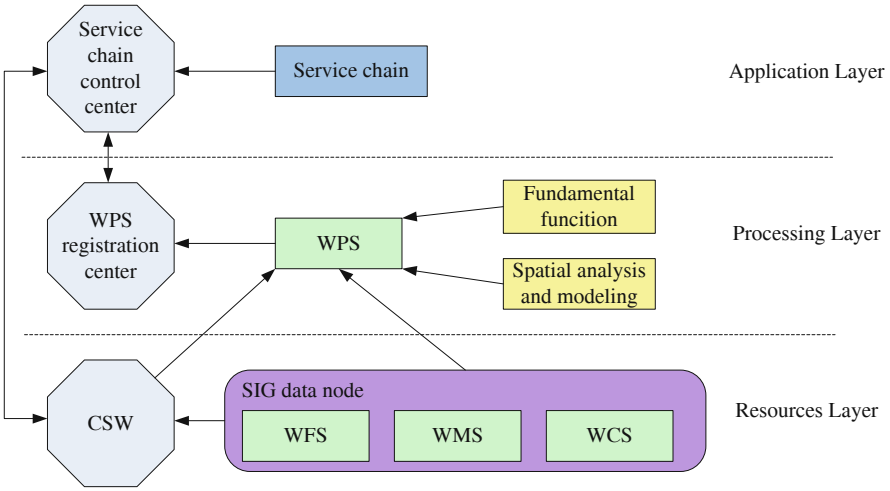

Fig. 2.10 Framework of service chain

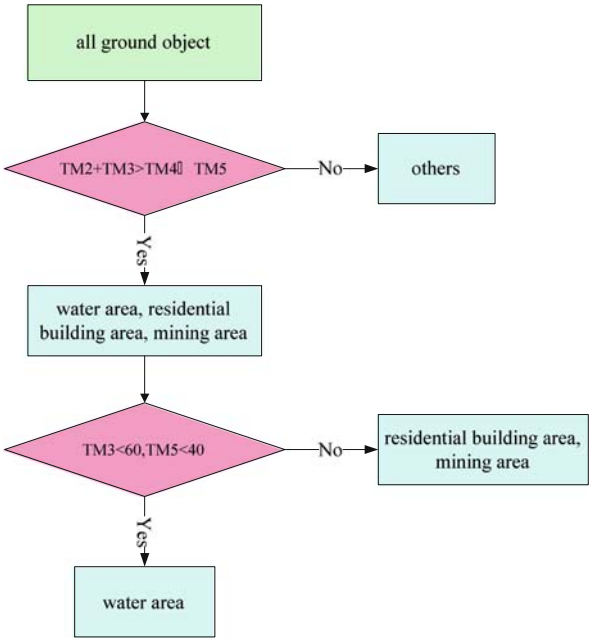

Fig. 2.11 Decision tree

org/standards/wps) (WPS). The WPS interface specifies three operations that can be requested by a client and performed by all WPS servers. Those operations are GetCapabilities, DescribeProcess, and Execute. These operations have many similarities to those in OGC Web Services such as WMS, WFS, and WCS.

In this test, all geospatial data are provided from distributed systems by the SIG data node. Processing functions are packaged as Web services, and can be discovered on line. For complex processing, which involves applying a chain of Web services-based geospatial processing functions, WPS is employed to link data and functions together based on a decision tree (see Fig. 2.11). A Web-based prototype system for extracting water information from TM remote sensing images was developed (see Fig. 2.12). This system shows the efficiency of the SIG data node and the ability to organize and execute a designed services chain. Furthermore, it has the advantages of being independent of platforms and program languages, and complying with OGC specifications.

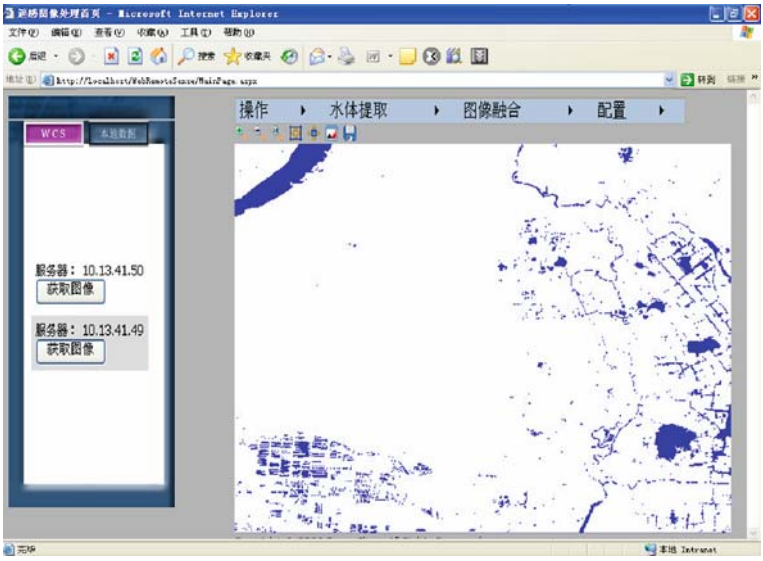

Fig. 2.12 Application interface

The experimental results indicate that these service nodes can successfully respond to requests compliant with OGC standards and return specific spatial data under different network environments. Any OGC WCS1.0-compliant geospatial information analysis system's client can access the data nodes distributed at different locations and locally retrieve geospatial data and complete comprehensive analysis. It shows that the SIG data node designed here can share distributed data. The Interoperable Access between the Windows data node and the Linux data node also shows that data can be shared under a heterogeneous operation environment.

2.6 Conclusions

Construction of the spatial information service node is fundamental for developing an infrastructure. This chapter discussed the design of a SIG data service node

based on NWGISS, which implemented WCS, WMS, CS/W, tools for transforming relevant data format, and several node instances. To satisfy demands from different applications, AXIS packages OGC Web Services and provides services to clients in three different ways. Several major OGC specifications are implemented. Concurrent with the development of OGC specifications, adding additional OGC specifications and managing of larger volumes of data will be investigated.

Acknowledgments The research presented in this paper was supported by the National High Technology Research and Development Program of China (863 Program), grant number 2003AA135118, 2008AA12Z2475774 and 2009AA12Z212.

References

Di, L., Chen, A., Yang, W., and Zhao, P. 2003. The Integration of Grid Technology with OGC Web Services (OWS) in NWGISS for NASA EOS Data[C]. Seattle, WA, USA: Proceedings of the Eighth Global Grid Forum (GGF8), June 24–27.

Di, L., Suresh, R., Doan, K., Ilg, D., and McDonald, K. 1999. DIAL-an Interoperable Web-Based Scientific Data Server. In M. Goodchild, M. Egenhofer, R. Fegeas, and C. Kottman (eds), Interoperating GeographicInformation Systems, Section 4. System.

Di, L., Yang, W., Deng, M., Deng, D., and McDonald, K., 2001. The Prototypical NASA HDF-EOS Web GIS Software Suite (NWGISS). http://laits.gmu.edu/Papers/NWGISS.htm.

Di, L., Yang, W., Deng, M., Deng, D., and McDonald, K., 2002. Interoperable Access of Remote Sensing Data Through NWGISS Geoscience and Remote Sensing Symposium, IGARSS'02, 1, 255–257.

Foster, I., Kesselman, C., and Tuecke, S., 2001. The Anatomy of the Grid-Enabling Scalable Virtual Organizations. Int. J. Supercomput. Appl. 15, 1365–1394.

Guo, D., Chen, H., and Luo, X., 2004. Resource Information Management of Spatial Information Grid. GCC 2003, LNCS 3033, pp. 240–243.

HDF-EOS4(HE4), http://hdfeos.gsfc.nasa.gov/hdfeos/he4.cfm.

ISO/TC 211 Geographic Information/Geomatics. http://www.isotc211.org/Outreach/Overview/Overview.htm.

ISO/TC 211.19115 Geographic Information – Metadata.http://www.isotc211.org/Outreach/Overview/Factsheet_19115.pdf.

ISO/TC 211.19119 Geographic Information – Services.http://www.isotc211.org/outreach/overview/Factsheet_19119.pdf.

King, M.D., (eds), 1999. EOS Science Plan. National Aeronautics and Space Administration, Washington D.C., NP-1998-12-069-GSFC.

King, M.D., Closs, J., Spangler, S., and Greenstone, R., (eds), 2003. EOS Data Products Handbook. National Aeronautics and Space Administration, Washington D.C., NP-2003-4-544-GSFC.

LAITS. HDF-EOS GIS translators, http://laits.gmu.edu/DownloadInterface.html.

LAITS. NASA HDF-EOS Web GIS Software Suite (NWGISS). http://nwgiss.laits.gmu.edu/introduction.htm.

Luo, Y., Wang, X., and Xu, Z., 2004. Spatial Information Grid – An Agent Framework. GCC 2003, LNCS 3032, pp. 624–628.

Luo, Y., Xue, Y., Wu, C., Hu, Y., Guo, J., Wan, W., Zhang, L., Cai, G., Zhang, S., and Wang, Z., 2006. A Remote Sensing Application Workflow and Its Implementation in Remote Sensing Service Grid Node. ICCS 2006, Part I, LNCS 3991, pp. 292–299.

McDonald, K. and Di, L., 2003. Serving NASA EOS Data to the GIS Community Through the OGC-standard Based NWGISS System. In Proceedings of 2003 Asia GIS Conference. Asia GIS Society. Oct 16–18, Wuhan, China. 13p.

National Imagery Transmission Format Version 2.1, National Imagery and Mapping Agency (NIMA), Bethesda, Maryland.

OGC.Catalogue Services. http://www.opengeospatial.org/standards/cat
OGC.Geography Markup Language.http://www.opengeospatial.org/standards/gml
OGC. OpenGIS Specifications. http://www.opengeospatial.org/standards/as
OGC.Web Coverage Service, http://www.opengeospatial.org/standards/wcs
OGC.Web Feature Service. http://www.opengeospatial.org/standards/wfs
OGC.Web Map Service.http://www.opengeospatial.org/standards/wms
OGC.Web Processing Service, Version 1.0.0. http://www.opengeospatial.org/standards/wps
Ren, Y., Fang, C., Chen, H., and Luo, X., 2004. Research on the Application of Multi-agent Technology to Spatial Information Grid. GCC 2003, LNCS 3032, pp. 560–567.
Ritter, N. and Ruth, M. GeoTIFF Format Specification GeoTIFF Revision 1.0, http://remotesensing.org/geotiff/spec/geotiffhome.html.
Shao, Z. and Li, D., 2005. Spatial Information Multi-grid for Data Mining. ADMA 2005, LNAI 3584, pp. 777–784.
Tang, Y., Chen, L., He, K., and Jing, N., 2004. A Study on System Framwork and Key Issues of Spatial Information Grid. Int J Remote Sens, 8(5): 425–433.
Tang, Y. and Jing, N., 2004. Research on System Architecture and Service Composition of Spatial Information Grid. GCC 2004 Workshops, LNCS 3252, pp. 123–131.
The Apache Software Foundation. Axis Architecture Guide. http://ws.apache.org/axis/java/architecture-guide.html.
Yang, W. and Di, L. Serving NASA HDF-EOS Data through NWGISS Coverage Server. http://laits.gmu.edu/Papers/NWGISS2.htm.

Chapter 3
Data Integration Support to the Coordinated Enhanced Observing Period Project (CEOP)

Kenneth R. McDonald, Yonsook Enloe, Liping Di, and Daniel Holloway

3.1 Introduction

While it is certainly true that many Earth science research and applications projects suffer from the lack of necessary environmental data and information products, an equally daunting challenge is the ability to effectively access, analyze, compare and integrate those products that do exist. This is especially true when the information products come from different sources with varying spatial and temporal resolutions and are processed and made available from different data systems with their own particular services, characteristics and conventions. Addressing this challenge requires a true partnership and the combined efforts of experts in the particular science or application discipline with those who have expertise in data systems capabilities and information technology.

An illustrative example of such a partnership is seen in the efforts of the Committee on Earth Observation Satellites (CEOS) Working Group on Information Systems and Services (WGISS) to provide data integration support to an international science program, the Coordinated Enhanced Observing Period (CEOP). This collaboration was initiated by the Lead Scientist for CEOP who was familiar with WGISS and requested that WGISS applies its experience and capabilities to assist the CEOP scientists with the access and integration of data and information products relevant to their study of the Earth's water and energy cycle. This partnership has been in place for over five years and the joint effort has just recently been completed. In the sections that follow, the technical details of the project and the resulting capabilities are described, along with our observations on the elements and characteristics of a successful IT application.

K.R. McDonald (✉)
NOAA/NESDIS/OSD/TPIO, Silver Spring, MD 20910 (formerly with NASA/GSFC, Greenbelt, MD 20771), USA
e-mail: Kenneth.Mcdonald@noaa.gov

L. Di, H.K. Ramapriyan (eds.), *Standard-Based Data and Information Systems for Earth Observation*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-540-88264-0_3,

3.2 Background

CEOS was established under the auspices of the Economic Summit of Industrialized Nations in 1984 in response to a recommendation from a panel of experts in remote sensing within the Working Group on Growth, Technology and Employment (CEOS, 2009). The panel recognized the collective value of the world's Earth remote sensing capabilities and the advantages that would be gained by the coordination of civil Earth observing satellite missions. By cooperating in mission planning and the development of compatible data products, applications, services and policies, the national space programs would maximize the benefits of their individual investments and be able to better address the environmental challenges of the entire international community. CEOS was to serve as the focal point for this international coordination and to provide the forum for the change of policy and technical information.

The members of CEOS are governmental organizations that are international or national in nature and are responsible for a civil space-borne Earth observation program that is currently in operation or in an advanced stage of system development. CEOS also has established Associate Members that are similar governmental organizations with a civil space-segment activity in an early stage of system development or those with a significant ground-segment activity that supports CEOS objectives. Associate Members may also be existing satellite coordination group and scientific or governmental bodies that are international in nature and have a significant programmatic activity that likewise is aligned with the goals of CEOS.

To accomplish its objectives, CEOS has created three working groups with members drawn from the each CEOS agencies with expertise in the particular topic area. The Working Group on Calibration/Validation ensures the accuracy and quality of Earth observation data and products through the international exchange of technical information and documentation, joint experiments and the sharing of facilities, expertise and resources (WGCV, 2009). The Working Group on Education, Training and Capacity Building (WGEdu) is focused on the coordination and partnership of CEOS members in providing education and training in Earth observation techniques, data analysis and interpretation, and applications, particularly in developing countries (WGEdu, 2009). The Working Group on Information Systems and Services (WGISS) promotes collaboration in the development of systems and services, based on international standards, which manage and supply Earth observation data and information from CEOS agency missions (WGISS, 2009).

All of the working groups provide a forum for exchange of information among members but for WGISS this has been especially important and beneficial. Over the past twenty years, the advances and rapid evolution in information technology have provided many opportunities for WGISS members to share their experience and expertise in the use of advanced IT systems to provide archive, discovery, access, visualization, analysis and utilization services to their agencies' data and information resources. Together, the members explored technologies such as web services, distributed systems, GIS capabilities, Grid services and sensor webs and

have tracked and contributed to the development of data, metadata and systems standards. The typical mechanisms to support these activities were member reports, technology demonstrations and the development of prototype capabilities.

While the exchange of information from these activities did benefit the WGISS members in the development of their individual agency capabilities, the idea emerged that a better way of evaluating the technologies and enhancing the overall WGISS capabilities would be to apply them to real requirements of the Earth science and applications user communities. The concept that emerged from this idea was that WGISS would develop a portfolio of its capabilities and expertise and share this with members of the Earth science and applications communities who were engaged in international programs. If that community has requirements that could be met by WGISS and was interested in collaborating, WGISS would establish a project to address the common objectives. The projects were called WGISS Test Facilities (WTF) reflecting the fact that WGISS relied on the best efforts of its member agencies and could not assume a long-term, operational responsibility (Best, 2000; Doyle, 2000). However, by working together, WGISS could develop advanced capabilities that would be useful and potentially integrated within the science or application program while at the same time enhancing the overall WGISS portfolio.

The Coordinated Enhanced Observing Period (CEOP) was created as an initiative of the Global Energy and Water Cycle Experiment (GEWEX) of the World Climate Research Programme (WCRP) with the goal of being able to understand and predict continental to local-scale hydroclimates with application to water resources. In 2007, the CEOP initiative was merged into the *Coordinated Energy and Water Cycle Observations Project (CEOP)*, which oversees all GEWEX Hydroclimate projects (Roads et al., 2007; GEWEX, 2009). A primary objective of CEOP was to produce consistent research quality data sets of the Earth's energy budget and water cycle and their variability and trends on inter-annual to decadal time scales. These data sets were to be made available to support the research and analysis goals of the CEOP and general Earth science community (CEOP, 2009).

To meet its objectives, CEOP undertook the assemblage of a diverse collection of in-situ data, time series and gridded model output and remotely sensed Earth observation data for 35 reference sites around the globe during a series of enhanced observing periods (EOPs). The CEOP program identified a series of products that would be of interest to its science community from a wide range of providers and established agreements for those providers to transfer their products to three designated CEOP data archives. The in-situ data were from field stations operated by local researchers and were sent to the National Center for Atmospheric Research (NCAR) for quality checking, archive and distribution. For each observing period, subsets of imagery and derived products from relevant Earth observing satellite missions for the reference sites and surrounding regions were provided by participating space agencies (Japanese Aerospace Exploration Agency (JAXA), National Aeronautics and Space Agency (NASA), etc.) and delivered to the University of Tokyo CEOP Satellite Data Integration Center (CSDIC) for archive, distribution and data integration services. Finally, for each reference site and observing period, gridded and

time-series output from Numerical Weather Prediction models from twelve institutes around the globe were assembled for archive and distribution at the World Data Center (WDC) for Climate hosted at the Max Planck Institute for Meteorology in Hamburg, Germany (Burford et al., 2007).

3.3 Data Integration Challenges

The collection of the rich set of CEOP data and products at the three archive facilities constitutes a valuable resource for the water and energy cycle research community, but major challenges remained in enabling its use. Within each data type, the original data and products were produced with differing processes and guidelines, have different spatial and temporal resolutions, were output in various formats, and are made available through different interfaces. The differences are even greater across data types and such heterogeneity is a major barrier to data integration and thus to the scientific study that require the data. The true value of these data resources are only realized when they can be inter-compared and combined by the research scientists. Then the accuracy and temporal coverage of the field data and the spatial coverage and regional context provided by the satellite products can be used to evaluate and refine the understanding of the physical processes represented by the models.

The CEOP approach that collects all of the data and information products of a particular type at a single location is one way that begins to address part of the problem. The archives for each of the three data types can impose certain standards on their respective sets of data providers that apply to their data submissions. As experts in the particular data types, they can also apply certain translations or transformations to the archived data and can redistribute those products following their own standards and guidelines. Further, they can create their own databases of metadata to catalog the data and information products. Each of the CEOP archives took such steps to enable data integration in the support of their particular data type collections.

Integrating data from the multiple CEOP archives presented an additional set of challenges, stemming in large part by the tools and associated interfaces that are used by different segments of the community to access the different data types. In the oceanographic and meteorological communities, the data access and transport capabilities of the Open-source Project for a Network Data Access Protocol (OPeNDAP) (Cornillon et al., 2003; OPeNDAP, 2009) are widely used and are an integral part of numerous research programs. The OPeNDAP is free software implementing the Data Access Protocol (DAP) version 2.0, which is a NASA Earth science data access standard (Gallagher et al., 2007). For the land science and geographic information systems (GIS) communities a growing number of data providers are serving their products via systems based on the specifications of the Open Geospatial Consortium (OGC), especially the Web Coverage Service (WCS) specification (Whiteside and Evans, 2008). This is especially true for remotely sensed imagery and gridded products provided by the space communities Earth observation satellites (Di, 2006; Di

and McDonald, 2006). Each of the two mechanisms–OpeNDAP and OGC– represent a major advance in enabling the integration of data and information products of a particular type but do impose a barrier to the integration and inter-comparison across data types.

3.4 WGISS-CEOP Partnership

Due to the broad scope of water and energy cycle research, the CEOP science community was interested in the full range of data products held at the three archive centers. Several CEOP scientists used OPeNDAP clients to access data from OPeNDAP servers but they were also interested in data served according to OGC specifications. WGISS members had extensive experience with both protocols from data management system programs within their home agencies and through prototype activities conducted as WGISS tasks. Through a series of interactions at their respective meetings, the WGISS members gained a better understanding of the data integration challenges of the CEOP program and the CEOP participants learned about the capabilities of WGISS. The clear needs of the CEOP scientists and their interest in the services that could be provided by the WGISS systems experts led to the formation of a WGISS Test Facility for CEOP (WTF-CEOP).

A fundamental purpose of a WTF is to explore various technology options and different methods to meet specified data and information systems requirements. In the case of the WTF-CEOP, the WGISS team took two different but complementary approaches. One effort that was led by JAXA was to work with each of the three CEOP archives to assist them in providing their data via an OPeNDAP server. Once the OPeNDAP servers were installed, configured and tested the collections at the three archives could be accessed by a single OPeNDAP client or application. The primary WGISS contribution was to provide training on the OPeNDAP capabilities and to assist the teams at the three archives to implement these new sets of services.

3.4.1 JAXA Contribution

As previously mentioned, through the initial interactions between WGISS and CEOP it was learned that many of the CEOP scientists were familiar with using OPeNDAP client tools to access and used environmental data. At that time, CEOP was beginning to aggregate the data of particular types (in situ, satellite and model outputs) from the first of the enhanced observing periods at the three designated archive facilities. A logical and manageable first step toward data integration was to work with the developers at the three archives to build the capabilities to provide their data via OPeNDAP servers. In a number of cases, the archive personnel were not experienced in the software and tools that are available in the OPeNDAP community. The JAXA team provided the necessary coordination and training support to the archive developers as their contribution to the WTF-CEOP.

3.4.2 NASA Development

The JAXA approach was a major step forward in enabling the integration of data products that were available at the three CEOP archives and represents a very viable solution for the CEOP archive model. However, it also illustrates that there are additional challenges in data access and integration that are not addressed by such an archive model. While that model does ensure that the CEOP products are available from the established CEOP archives that are sustained partners in the CEOP program, it carries the implicit assumption that the archives hold all of the products that are of interest to the CEOP scientists and the most complete and current versions of those products. That will never be the case because of the tremendous set of environmental data resources and the dynamic nature of the products due to the reprocessing of satellite data products and model reanalysis.

Just as it is impossible to expect that all relevant data for a particular scientific endeavor can be aggregated at a few central facilities, it is also unreasonable to expect that all sources of such products will make them available via the same standard interfaces and protocols. However, developing mechanisms that can handle a small number of interfaces or protocols would be a major advance in the support of data access and integration and this was the goal of the NASA WTF-CEOP team.

3.5 NASA's Data Integration Gateway

A growing number of satellite data providers are serving their data following the Open Geospatial Consortium's (OGC) Web Coverage Server (WCS) specification. Representatives from the CEOP science community were interested in having access to several NASA WCS data collections but wanted to continue using their OPeNDAP clients. The NASA WTF-CEOP team developed a concept for a prototype that would provide a gateway between the OPENDAP and WCS protocols. The gateway allows a user to use an OPeNDAP enabled client to access satellite data held at a WCS server with services such as subsetting, reprojection, mosaicking and time series support. The project would develop a data handler for the OPeNDAP server that enabled serving of data obtained from a WCS server. The aim was to allow access to the data by analysis clients that have OPeNDAP support but not WCS support. For its part, the WCS server enables the serving of high resolution satellite swath data suitably reprojected to the WCS spatial reference system in a gridded form intelligible to the OPenDAP client. A useful byproduct of the gateway architecture also allows third parties to provide OPeNDAP interfaces to data they do not actually hold but simply access remotely through WCS.

Since WGISS does not have dedicated funding but rather relies on the contributions of its members, the first task of the NASA team was to develop a complete proposal and secure funding to develop the prototype. The team members were from NASA Goddard Space Flight Center, SGT, Inc., George Mason University's Center for Spatial Information Science and Systems, and OPeNDAP, Inc. Together,

and with the strong support of the CEOP community, the team won a three-year research award from NASA's Advancing Collaborative Connections for Earth-Sun System Science Program (ACCESS) to develop a CEOP Satellite Data Server.

The CEOP Satellite Data Server was initially designed to include a gateway (middleware) between a WCS server for NASA satellite data and the OPENDAP servers. However, the re-architecture of the new version of the OPeNDAP server, Hyrax, allowed the key gateway functionality to be implemented instead as a "format handler" in the Hyrax's Back-end Server (BES). Figure 3.1 shows the eventual high level design of the CEOP Satellite Data Server. The handler can be configured during the installation of the Hyrax server. In addition, significant customization of the WCS server implementation and configuration was necessary as well and forms the main basis for the lessons.

Two separate WCS servers were implemented, though sharing a substantial codebase. A WCS server was implemented to serve daily global coverages of an Atmospheric Infrared Sounder (AIRS) Level 2 Standard Retrieval product (aka AIRX2RET) with actual CEOP requests focused on CEOP observation sites, essentially 250x250 km squares around CEOP reference sites. Although in theory, this server could respond to general, arbitrary user WCS requests, its primary purpose was to act as the back end to the CEOP Satellite Data Server, performing essential reprojection and mosaicking of AIRS Level 2 data, Thus, the 250x250 km CEOP reference site squares might be thought of as virtual products, generated on the fly. The global daily coverages presented to the client are actually the virtual products, as they are not archived in the Goddard Earth Sciences Data and Information

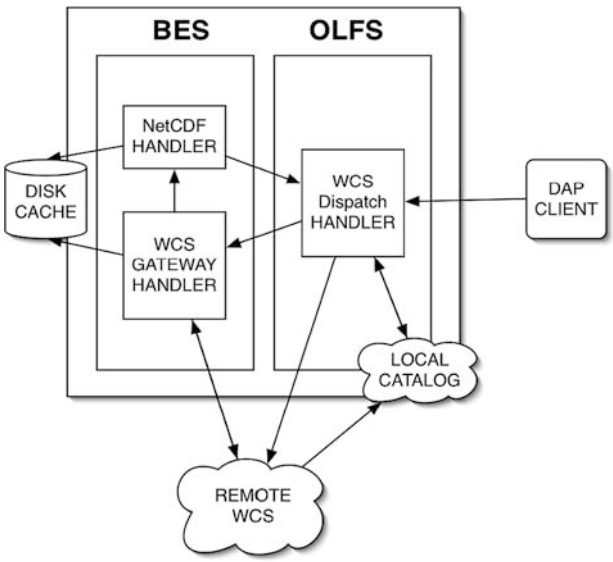

Fig. 3.1 High-level conceptual design of the OPeNDAP/WCS server

Services Center (GES DISC). The WCS server generates these coverages from individual 6-min granule files physically stored in the GES DISC's operational archive. The second WCS server was implemented within George Mason University's environment to serve MODIS data. In this implementation, individual granules were mapped to coverages.

The basic process of fulfilling a request is:

1. The client submits an OPeNDAP request to the CEOP Satellite Data Server, which is handled by its front end, the OPeNDAP Light Front End Service (OLFS).
2. the OLFS interacts with local catalog to identify the data source as WCS;
3. the OLFS instructs its BES to set container type to WCS and passes identifying information about the data to be retrieved;
4. BES formulates a WCS request to the WCS server;
5. BES stores the WCS response to local cache;
6. BES uses the NetCDF format handler to process cached file to satisfy the Data Access Protocol (DAP) request; and
7. Subsequent DAP requests operate against local cache.

Several times during the development of the prototype, CEOP scientists were recruited to test the prototype and provide feedback on the usability of the current state of the prototype and suggestions for improvement of the prototype. The feedback provided was key to driving the prototype development. Feedback on what data was accessed and over what time periods and the pattern of access was key. Feedback on what the user expected to see and the types of information needed by the user to understand that data that he was accessing was also important.

3.6 Outcomes and Lessons Learned

In addition to the data integration services that have been developed and put into place by the WTF-CEOP activities, a number of programmatic and technical lessons have emerged over the course of the project. The most important are related to the strong partnership that was formed between the CEOP project and the WGISS participants. As previously mentioned, this partnership very likely played a major role in WGISS members securing the support of their respective agencies to develop the WTF and the CEOP science input on requirements definition and their participation in user testing were essential in the development of the prototypes. In the case of the NASA CEOP Satellite Data Server, the user input for the project revealed several challenges not normally faced by WCS servers. A key revelation was the expectation that many of the users in question were likely to want to analyze long time series of data for very small geographic areas. This contrasts with most WCS access patterns, which tend to concern short time periods (or even neglect time entirely.) This in turn raised a number of challenges that rippled throughout many aspects of the project, including the WCS server design and implementation. Finally, user

questions about the content of what they were actually seeing pointed out the need to consider quality screening and provenance in the WCS server.

Another lesson from the WTF-CEOP experience is that there are multiple ways to attack the challenges of data integration and each will have its advantages and disadvantages. The CEOP approach of aggregating collections of a common type at several central archives does ensure the availability of a high-quality set of core data sets and did facilitate the JAXA initiative to support the implementation of OPeNDAP servers at each of the archive facilities. The collections however are fixed and somewhat limited. By providing similar access capabilities to data directly from the source, the NASA CEOP Data Server provides access to a wider set of data collections, including the most recently reprocessed versions, but subject to the varying policies and practices of the source providers. Together, the capabilities developed by JAXA and NASA provide a broad and complementary set of data integration services.

The WTF-CEOP by its nature and intent was a technology research project to develop and apply new capabilities in the integration of heterogeneous data types in support of a particular set of Earth science studies. As might be expected with this type of investigation, the actual implementation did evolve from the original design concept over the course of the project. The functionality of the gateway ended up being implemented within the Hyrax OPeNDAP server and while the transition of the CEOP Satellite Data Server to an operational capability for the CEOP program has not yet been realized this represents a very significant outcome. Because the Hyrax server maintenance is supported by the OPeNDAP, Inc, a non-profit organization, implementing the key gateway functionality as a "format handler" in Hyrax allows this capability to live long after the ACCESS funding for the NASA activities. It is hoped that CEOP and other science and applications programs with similar data analysis requirements will be able to benefit from both the WTF-CEOP lessons learned and the data integration services it has advanced.

Acknowledgements Drs. Christopher Lynnes and Wenli Yang from the NASA Goddard Space Flight Center and George Mason University, respectively, have played major roles in the development of the NASA CEOP Satellite Data Server in the later stages of the project and their contributions are hereby acknowledged.

References

Bset, C., 2000. A Possible WGISS Test Facility. http://www.intl-interfaces.com/wtf-gofc/2000.11.27_WTF-GOFC_Tech_Team_Greenbelt/Clive_Best_WGISS_Test_Facility.pdf (Accessed on May 25, 2009).

Burford, B., Ochiai, O., Enloe, Y. and McDonald, K., 2007. Distributed Data Integration Services Provided by the WGISS Test Facility for CEOP. *Journal of the Meteorological Society of Japan*, 85A:519–527.

CEOP, 2009. CEOP homepage. http://monsoon.t.u-tokyo.ac.jp/ceop2/ (Accessed on May 25, 2009).

CEOS, 2009. CEOS Background, http://www.ceos.org/index.php?option=com_content&view=category&layout=blog&id=25&Itemid=73 (accessed on May 25, 2009).

Cornillon, P., Gallagher, J., and Sgouros, T., 2003, OPeNDAP: Accessing Data in a Distributed, Heterogeneous Environment, *Data Science Journal*, 2:164–174.

Di, L., 2006. "The Open GIS Web Service Specifications for Interoperable Access and Services of NASA EOS Data." In Qu, J. et al. eds, *Earth Science Satellite Remote Sensing*. Springer-Verlag, New York, pp. 254–268.

Di, L. and McDonald, K., 2006. "The NASA HDF-EOS Web GIS Software Suite (NWGISS)." In Qu, J. et al. eds, *Earth Science Satellite Remote Sensing*. Springer-Verlag, New York, pp. 282–292.

Doyle, A., 2000. WGISS Test Facility Target Vision/Demo Synthesis. http://www.intl-interfaces.com/2000.09.11_ceos_meetings/2000.09.13_WGISS_Test_Facility_Presentation.ppt (Accessed on May 25, 2009).

Gallagher, J., Potter, N., Sgouros, T., Hankin, S. and Flierl, G., 2007. The Data Access Protocol—DAP 2.0. NASA ESE-RFC-004.1.1. http://www.opendap.org/pdf/ESE-RFC-004v1.1.pdf (Accessed on May 25, 2009).

GEWEX, 2009. GEWEX-CEOP homepage. http://www.gewex.org/ceop.htm (Accessed on May 25, 2009).

Roads, J., Benedict, S., Koike, T., Lawford, R. and Sorooshian, S., 2007. Towards a new Coor-dinated Energy and Water-Cycle Observations Project (CEOP): Integration of the Coordinated Enhanced Observing Period (formerly known as 'CEOP'*) and the GEWEX Hydrometeorology Panel (GHP). http://www.gewex.org/GHP-CEOPmerger-whitepaper.pdf (Accessed on May 25, 2009).

WGCV, 2009. CEOS Working Group on Calibration and Validation. http://www.ceos.org/index.php?option=com_content&view=category&layout=blog&id=75& Itemid=113 (Accessed on May 25, 2009)

WGEdu, 2009. CEOS Working Group on Training and Education. http://www.ceos.org/index.php?option=com_content&view=category&layout=blog&id=104&Itemid=153 (Accessed on May 25, 2009).

WGISS, 2009. CEOS Working Group on Information Systems and Services. http://wgiss.ceos.org/ (Accessed on May 25, 2009).

Whiteside, A. and Evans, J., 2008. eds, Web Coverage Service (WCS) Implementation Standard. OGC^{TM} Document: 07-067r5. http://portal.opengeospatial.org/files/?artifact_id=27297 (Accessed on May 25, 2009).

Chapter 4
Progress in OGC Web Services Interoperability Development

George Percivall

4.1 OGC Vision for Geospatial Interoperability

Decision makers in business and government have historically depended on geomatics experts when they have sought to benefit from Earth observation systems. Similarly, scientists in fields other than geomatics have had to either learn about geomatics or team with geomatics experts to benefit from these systems. Fortunately, as Earth observation technologies and markets have progressed, standards have steadily advanced, which, along with other benefits described below, allows geomatics experts to establish reusable services for routine decision-making.

The Open Geospatial Consortium, Inc. (OGC) is the organization that has most prominently and successfully created and promoted Web services standards for geoprocessing and geospatial decision support. OGC's vision is the realization of the full societal, economic and scientific benefits of integrating electronic location resources into commercial and institutional processes worldwide. Is support of this vision, OGC's mission is to serve as a global forum for the collaboration of developers and users of spatial data products and services, and to advance the development of international standards for geospatial interoperability (http://www.opengeospatial.org/ogc/vision)

4.1.1 OGC Overview

The OGC is a not for profit, international voluntary consensus standards organization founded in 1994. The core mission of the OGC is to develop standards that enable interoperability and seamless integration of spatial information, processing software, and spatial services. Spatial information and processing encompass geographic information systems (GIS), remote sensing, surveying and mapping, navigation, location-based services, access to spatial databases, sensor webs, and other spatial technologies and information sources.

G. Percivall (✉)
Open Geospatial Consortium, Inc., Herndon, VA 20170-4819, USA
e-mail: gpercivall@opengeospatial.org

L. Di, H.K. Ramapriyan (eds.), *Standard-Based Data and Information Systems for Earth Observation*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-540-88264-0_4,

Spatial information and processing play an important role in business, government, and research applications and workflows. However, the benefits of using spatial information and services are often limited by the inability to effectively share information between different vendors' solutions and different types of systems. In the OGC's consensus process, over 360 government, private sector, and academic organizations cooperatively define, develop, test, document, validate, and approve interface and encoding standards that overcome the interoperability problems.

The OGC baseline of adopted standards includes these implementation specifications (http://www.opengeospatial.org/ogc/standards):

- Web Map Service (WMS)
- Web Feature Service (WFS)
- Web Coverage Service (WCS)
- Catalogue Service for the Web (CSW)
- Sensor Observation Service (SOS)
- Sensor Planning Service (SPS)
- Sensor Alert Service (SAS)
- Geography Markup Language (GML)
- Web Map Context
- KML

The OGC standards baseline allows for service-oriented architecture as shown in Fig. 4.1.

Specific interoperability requirements are brought into the OGC process by government agencies, vendors, and universities, and by integrators working on behalf of their customers. In OGC testbeds, interoperability experiments, and pilot projects, sponsoring organizations pool their interoperability requirements and arrange incentives for technology providers to work together to make their systems work together. Sometimes technology developers submit interoperability requirements to meet anticipated needs in the marketplace.

4.1.1.1 Benefits for Technology and Content Providers

Competing technology and content providers collaborate in the OGC because they recognize that lack of interoperability is a bottleneck that slows market expansion. Interoperability enabled by open standards positions them to both compete more effectively in the marketplace and to seek new market opportunities. In the OGC, technology and content provider members:

- Position themselves early to influence definition of new open standards.
- Reduce costs through cooperative standards development with other OGC members.
- Shorten time to market by using OGC standards rather than custom interfaces.
- Can enter new markets and find new customers because of "plug and play".

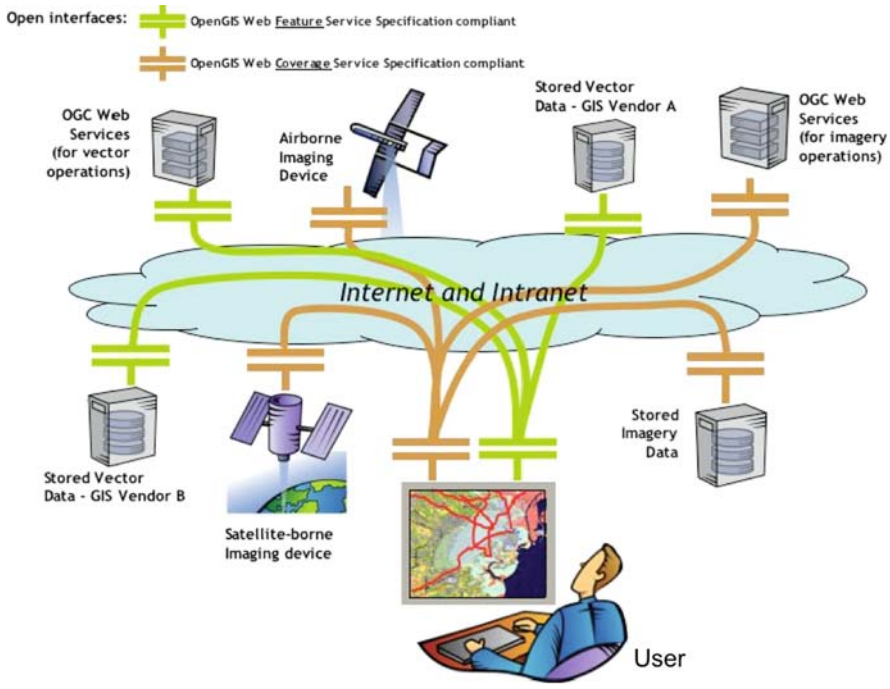

Fig. 4.1 Geospatial Web services based on open interfaces and encodings enable users and software to access diverse data and processing resources on the Web

- Have a convenient forum for discussion of industry issues and solve shared problems.
- Form customer relationships and business partnerships.
- Deliver solutions more quickly and at lower cost.
- Can mobilize a range of products across open interfaces, rather than performing resource intensive custom integration.
- Provide precise solutions to meet specific needs, solutions that plug and play.

The development of these standards does not require the member to give up any intellectual property or trade secrets. The use of open standards to connect components, applications, and content – allowing a white box view on the components' functionality and interrelationships without revealing implementation details – fulfills the industry requirement for protection of intellectual property as well as the user requirement for transparency. Such transparency supports both interoperability and the credibility of the enterprise, or federated solution.

4.1.1.2 Benefits for Technology Consumers

Technology consumer members can:

- Voice their interoperability needs directly to a broad and global industry, academic and government community. In this setting, vendors, integrators, and platform providers build interoperability interfaces far faster than is possible with traditional system integration contracting. And the benefits are shared globally.
- Be assured that reusability of software is achieved. This is often cited as the single greatest benefit anticipated from complying with standards or helping to establish them.
- Work with other users in the OGC process to demonstrate the need for and potential market appeal of new requirements for specifications.
- View OGC programs as a form of technology risk reduction. Small resource investments in the OGC industry consensus processes often result in industry willingness to address and then broadly implement OGC specifications in their products. Standards help users maximize the return on their current and future technology investments, while reducing the time and cost of customized integration.
- Mobilize new technology solutions quickly, and adapt easily to the rapidly changing information technology world, policy changes, and new and emerging requirements.
- Leverage existing investments in legacy content and applications. The use of standards provides a fulcrum to leverage IT investments and create liquidity. Standards provide a platform for realizing opportunities that would otherwise remain hidden.
- See the OGC process as a method for procurement reform. Users benefit by expressing their interoperability requirements first in the OGC specification process, then by adopting procurement language that calls for OGC specifications in the geospatial and location based services products to be considered for purchase and deployment.

4.1.1.3 Policy on Intellectual Property

The OGC offers its specifications free of charge to all, and adheres to a rigorous process to ensure that OGC standards remain free of royalties for use. The OGC strongly supports royalty-free standards, a position also taken by the World Wide Web Consortium and other prominent standards organizations. The OGC believes that standards consortia play a major role in maintaining a free and open Web, and that open geoprocessing standards are an important part of the free and open Web.

The sections below provide more detail on OGC Web Services (OWS) standards.

4.1.2 OGC Standards Overview

The OGC consensus process for defining, developing, and approving a standard generates a number of documents. These documents are typically first developed in testbeds and interoperability experiments managed in the OGC Interoperability

Program. Then the draft standards work their way through the approval process in the OGC Specification Program, which includes the OGC Technical Committee and the OGC Planning Committee. Approved OGC standards detail the interface or encoding structures that, when implemented, enable interoperability between systems.

Standard interfaces, protocols, and encodings enable different software and application products to communicate, whether they are running on the same computer or they are exchanging instructions across the Web. These standards also enable much easier integration of complex systems. Standards are necessary ingredients for designing and implementing "open architectures" and for "service oriented architectures (SOA)" that need to be accessed by various clients and server processes running on diverse and unknown computing platforms across the Web. Standards also reduce dependence on single point solutions, reduce risk, reduce software lifecycle costs, and create new business opportunities for technology providers. Providing "open" access to data and services on the Web and making it discoverable through a spatial catalog exposes those data and services to a much larger set of potential users, and thus increases the data's value.

Below we describe some of the adopted OpenGIS Implementation Standards and other documents that are most relevant to Earth observation. The full set of OGC adopted standards are available online (http://www.opengeospatial.org/ogc/standards).

- The *OpenGIS® Web Map Service (WMS) Interface Standard* supports the creation and display of registered and superimposed map-like views of information that come simultaneously from multiple remote and heterogeneous sources. The servers can be servers of raster or vector data, or even scanned maps. The maps are delivered to the browser or other Web-based viewing application in simple Web graphic image formats. The OpenGIS® Styled Layer Descriptor (SLD) Implementation Standard extends the WMS specification to allow user-defined symbolization of feature data.
- The *OpenGIS® Web Coverage Service (WCS) Interface Standard* interface allows client and/or application query and access to geospatial "coverages", such as imagery and digital elevation models. The result of the coverage query (the actual data) is made available to the client, service, or application. The WCS operations allow for access to imagery including subsetting requests in space, time and parameters.
- The *OpenGIS® Web Feature Service (WFS) Interface Standard* enables a client or service that implements the interface to retrieve and optionally update geospatial feature data from any server that implements the WFS interface. The WFS interface does not "care" how the feature data are stored. The interface is content and storage model independent. The result of a WFS query is typically returned as a GML document.
- The *OpenGIS® Web Map Context (WMC) Interface Standard* is a companion specification to the Web Map Service Standard and allows an application to store "state" information. The XML-encoded Context document includes information

about the WMS servers providing layers to the overall map, the bounding box and map projection shared by all the maps, and so forth. This is sufficient metadata for any client software to reproduce the composite map, and ancillary metadata used to annotate or describe the maps and their provenance for the benefit of human viewers.

- The *OpenGIS® Web Image Classification Service (WICS) Interface Standard* deals specifically with classification of digital images. This draft specification provides a web based interface to image classification services of any type. This specification does not specify a particular classification algorithm. The interface allows a client to request that the service perform a classification on a source image resulting in a grid coverage feature with the attributes being categories. This specification allows for clients to request classification from a variety of algorithms. This document is currently a Discussion Paper, which after further development may become an implementation specification.
- The *OpenGIS® Web Coordinate Transformation Service (WCTS) Interface Standard* defines an interface to request transformation of geospatial data from one coordinate reference system (CRS) to another. Geospatial data including imagery are often stored in different CRSs. For an application to use data stored in different CRSs, such data must be transformed or converted into the same CRS. This service inputs digital features or coverages in one CRS and outputs the same features in a different CRS. This document is currently an OGC Discussion Paper, which may become an implementation specification.
- The *OpenGIS® Catalog Services Web (CSW) Interface Standard* defines common interfaces to discover, browse, and query metadata about data, services, and other potential resources. Once substantial numbers of data sets and geospatial Web services have been registered in such catalogs with metadata that conforms to ISO/CD TS 19139 (XML schema implementation), users (and automated processes) will have a far greater ability to find data and services.
- The *OpenGIS® Geography Markup Language (GML) Encoding Standard* is an XML-based language for encoding geographic information to be transported over the Internet or other transport environments. GML encodes both the geometry and properties of objects that comprise geographic information. GML allows the data to be controlled in the client by the user who receives geometries and geographic features and customizes how the data is to be displayed. Profiles and application schemas of GML can be defined to meet the requirements of specific information communities. An example is the new OpenGIS GML in the JPEG 2000 Implementation Standard, which defines the means by which the GML can be used within JPEG 2000 images for geographic imagery.
- The *OpenGIS® CityGML Encoding Standard* is an open data model framework encoding standard for the storage and exchange of virtual 3D urban models. It is an application schema of GML3. Computer Aided Design (CAD)/geospatial convergence is necessary because the architecture/engineering/construction industry and other domains often need to use building information in the context of diverse geospatial information. The use of CityGML can be of significant relevance to

the IEEE GRSS research in urban settings. (See the section regarding current developments in CAD/Geospatial integration.)

– The *OpenGIS® Web Processing Service (WPS) Interface Standard* provides rules for standardizing inputs and outputs (requests and responses) for geospatial processing services. The standard also defines how a client can request the execution of a process, and how the output from the process is handled. It defines an interface that facilitates the publishing of geospatial processes and clients' discovery of and binding to those processes. The goal is to provide consistency among the various types of geoprocessing services, to ensure the success of complicated service chaining and workflows involving multiple types of services.
– The *OGC Reference Model* is a document that describes all the OpenGIS Specifications, how they work together, and how they work in various distributed computing environments. This is a good place to begin a study of OGC's technical baseline and standards framework (http://www.opengeospatial.org/standards/orm).

4.1.3 Sensor Web Enablement (SWE)

Sensor technology, computer technology, and network technology are advancing together while demand grows for ways to connect information systems with the real world. The OGC's Sensor Web Enablement (SWE) standards enable developers to make all types of sensors, transducers, and sensor data repositories discoverable, accessible, and useable via the Web.

SWE standards are developed and maintained by OGC members who participate in the OGC Technical Committee's Sensor Web Enablement Working Group. OGC members have reached agreement on the most of the issues involving digital communication about the complex location, motion, and optical parameters involved in satellite-borne imaging systems and photogrammetry as well as virtually all aspects of other types of sensors and sensor data. SWE standards offer developers:

- Open interfaces for sensor web applications
- "Hooks" for IEEE 1451, TML, CAP, WS-N, ASAP
- Imaging device interface support
- Sensor location tied to geospatial standards
- Fusion of sensor data with other spatial data
- Opportunity to participate in an open process to shape standards, in cooperation with IEEE and other standards organizations

Below are listed the main adopted OpenGIS Standards in the SWE framework:

– The *OpenGIS(R) Observations & Measurements (O&M) Encoding Standard* provides general models and XML encodings for observations and measurements.

- The *OpenGIS(R) Sensor Model Language (SensorML) Encoding Standard* provides standard models and XML Schema for describing the processes within sensor and observation processing systems.
- The *OpenGIS(R) Transducer Markup Language (TML) Encoding Standard* provides a Conceptual model and XML encoding for supporting real-time streaming observations and tasking commands from and to sensor systems.
- The *OpenGIS(R) Sensor Observation Service (SOS) Interface Standard* provides open interface for a web service to obtain observations and sensor and platform descriptions from one or more sensors.
- The *OpenGIS(R) Sensor Planning Service (SPS) Interface Standard* provides an open interface for a web service by which a client can 1) determine the feasibility of collecting data from one or more sensors or models and 2) submit collection requests.

SWE was one of the main focus areas in OGC's 2005 OGC Web Services 3 (OWS-3) testbed activity. Important progress was made in harmonizing the SWE services listed above with the IEEE 1451 standard for plug-and-play sensors.

Other SWE standards are under discussion or in various stages of development.

4.2 Current Developments

The interfaces described above have been implemented in hundreds of commercial products, custom systems, and open source applications. (See http://www.opengeospatial.org/resources/?page=products.) But much work remains, and many other specifications are working their way through OGC's processes. Below is a summary of the main technology domains in which there is ongoing specification development activity:

- *Geospatial Rights Management (GeoRM)* – Efforts to manage data ownership and data rights in the digital environment are of great interest to geospatial data providers who need to control access to their data and how it is used. This is of concern to data sellers, to organizations whose data is for internal use only, and to those whose data distribution follows a library model. GeoRM involves persistent management of a geospatial digital object under a set of rights and conditions. GeoRM is now an approved OGC reference model.
- *Access control* is a necessary complement to rights management. The adopted OpenGIS® Geospatial eXtensible Access Control Markup Language (GeoXACML) Encoding Standard defines a geo-specific extension to the XACML Policy Language 2.0 (eXtensible Access Control Markup Language (XACML) Version 2.0), as standardized by the Organization for the Advancement of Structured Information Standards (OASIS). GeoXACML is likely to become one of the standards registered in the GEOSS Standards and Interoperability Registry. It is a key technology for enabling institutional interoperability.

- *Geoprocessing Workflow* – Geoprocessing creates geospatial information specific to a user's specific decision-making needs. In some cases, a chain of OGC Web services is needed to produce the specific value-added products needed. OGC Web Services can be chained together using the OASIS Business Process Execution Language (BPEL) specification. For example, the OGC Web Coordinate Transformation Service (WCTS) and the OGC Web Image Classification Service can be "chained" into an integrated workflow.
- The Geo-Processing Workflow (GPW) thread in the sixth OGC Web Services testbed, OWS-6, aims to develop and demonstrate interoperability among geoprocesses through service chaining, workflow and web services, with emphasis on implementing security capabilities for OGC web services, including SWE services. Work in this thread builds on the results from previous testbeds, including authentication/authorization and Simple Object Adaptor Protocol (SOAP)/ Web Services Description Language (WSDL) recommendations. The workflow and security tasks involve three operational security environments: 1) internal to a single trusted domain; 2) between two trusted domains; and 3) between a trusted and non-trusted (or temporarily-trusted) domain.
- *Decision Support Services*: In the past "decision support systems" have been monolithic applications that helped managers find solutions to difficult management problems. With the advent of the Internet and distributed web services it is now possible to define decision support as the coordination of various services that transparently convert geospatial data from other communities into terms familiar to the user. A decision maker is able to sit down at a single workstation, identify any geospatial resource anywhere, access that resource, bring it into the user's operational context, and integrate it with other resources to support the decision process. (See section on SOAP and REST.) The focus for DSS in the OWS-6 testbed builds on portrayal, WMS Tiling, integrated clients, and 3D visualization and integration of the built environment and landscape.
- Data portrayal requirements are often complex and largely based on feature attribution as opposed to simply feature types. To address complex symbology requirements, participants in the OWS-6 testbed are exploring ways to integrate the OGC's OpenGIS® Styled Layer Descriptor (SLD) Standard, a profile of the OpenGIS® Web Map Service (WMS) Encoding Standard, with the ISO standard for portrayal (ISO 19117 v.2.0) and the International Hydrographic Office (IHO) S52 symbology for maritime features.
- *CAD/Geospatial Integration*: Through Building Information Models (BIM) supported by various software vendors' products, professionals in the architecture/engineering/construction industry and related domains seek to make it easy to integrate building information of all kinds, including diverse geospatial information. To standardize BIM, the International Alliance for Interoperability (IAI) has provided Industry Foundation Classes (IFC), consensus-based open standards that support communication of project lifecycle data about a building (http://www.iai-tech.org/products/ifc_specification). But IFC is essentially a transfer standard or batch file conversion standard, analogous to the Spatial Data Transfer Standard (SDTS) introduced into the GIS industry in 1992 by the US Geological Survey

(USGS). Like SDTS, IFC is cumbersome, and vendors have had little guidance or incentive to ensure that their implementations are consistent with other vendors' implementations. To address this problem, the OGC and the buildingSMART alliance (bSa) have joined forces to develop candidate BIM standards in the joint bSa-OGC Architecture / Engineering / Construction / Owner / Operator Testbed (AECOO-1).
– Another current development in CAD/Geospatial integration is the OGC's Web 3D Service (W3DS). W3DS specifies a Portrayal Service and is a relatively new OGC discussion paper. In its current form, it provides a 3D representation of geographic data. The advantages of using visualization-centric formats are that they support a wide range of features for controlling the visual appearance (e.g. textures, surface properties, animations, lighting, atmosphere) and that they can be more efficiently transmitted and encoded.

At every OGC meeting, new requirements for interoperability are discussed. If there is sufficient interest and resource, work begins and the participants report the results of their work at the next meetings. Initiatives such as Testbeds and Interoperability Experiments are planned and executed. The scope of the work keeps expanding, and so do the number of adopted specifications, the number of implementations, and the number of people who are benefiting from the implementations.

4.2.1 OWS Architecture

In the early 1990s, the OGC defined a vision for network-based geospatial computing. This vision has come to fruition using Web services. This section describes the vision and the OGC Web Services architecture.

The widespread application of computers and use of geographic information systems (GIS) have led to the increased analysis of geographic data within multiple disciplines. Through advances in information technology, society's reliance on such data is growing. Geographic datasets are increasingly being shared, exchanged, and used for purposes other than their producers' intended ones. GIS, remote sensing, automated mapping and facilities management (AM/FM), traffic analysis, geopositioning systems, and other technologies for Geographic Information (GI) have entered a period of radical integration.

Standards for geospatial interoperability provide a framework for developers to create software that enables users to access and process geographic data from a variety of sources across a generic computing interface within an open information technology environment. To elucidate:

- "a framework for developers" means that the International Standards are based on a comprehensive, common (i.e., formed by consensus for general use) plan for interoperable geoprocessing.

- "access and process" means that geodata users can query remote databases and control remote processing resources, and also take advantage of other distributed computing technologies such as software delivered to the user's local environment from a remote environment for temporary use.
- "from a variety of sources" means that users will have access to data acquired in a variety of ways and stored in a wide variety of relational and non-relational databases.
- "across a generic computing interface" means that standard interfaces provide reliable communication between otherwise disparate software resources that are equipped to use these interfaces.
- "within an open information technology environment" means that the standards enable geoprocessing to take place outside of the closed environment of monolithic GIS, remote sensing, and AM/FM systems that control and restrict database, user interface, network, and data manipulation functions.

4.2.1.1 OWS Fundamentals

The fundamental principles of the OGC Web Services (OWS) architecture include:

1. Service components are organized into multiple tiers.
    a. All components provide services, to clients and/or other components, and each component is usually called a service (with multiple implementations) or a server (each implementation).
    b. Services (or components) are loosely arranged in four tiers, from Clients to Application Services to Processing Services to Information Management Services, but un-needed tiers can be bypassed.
    c. Services can use other services within the same tier, and this is common in the Processing Services tier.
    d. Servers can operate on (tightly bound) data stored in that server and/or on (loosely bound) data retrieved from another server.
2. Collaboration of services produces user-specific results.
    a. All services are self-describing, supporting dynamic (just-in-time) connection binding of services supporting publish-find-bind.
    b. Services can be chained with other services and often are chained, either transparently (defined and controlled by the client), translucently (predefined but visible to the client), and opaquely (predefined and not visible to client), see Subclause 7.3.5 of [ISO 19119]
    c. Services are provided to facilitate defining and executing chains of services.
3. Services communication uses open Internet standards.
    a. Communication between components uses standard World Wide Web (WWW) protocols, namely HTTP GET, HTTP POST, and SOAP.

b. Specific server operations are addressed using Uniform Resource Locators (URLs).
c. Multipurpose Internet Mail Extensions (MIME) types are used to identify data transfer formats.
d. Data transferred is often encoded using the Extensible Markup Language (XML), with the contents and format specified using XML Schemas.

4. Service interfaces use open standards and are relatively simple.

a. OGC web service interfaces are coarse-grained, providing only a few static operations per service.
b. Service operations are normally stateless, not requiring servers to retain interface state between operations.
c. One server can implement multiple service interfaces whenever useful.
d. Standard XML-based data encoding languages are specified for use in data transfers.

5. Server and client implementations are not constrained.

a. Services are implemented by software executing on general purpose computers connected to the Internet. The architecture is hardware and software vendor neutral.
b. The same and cooperating services can be implemented by servers that are owned and operated by independent organizations.
c. Many services are implemented by standards-based Commercial Off The Shelf (COTS) software.

4.2.1.2 OWS Services Tiers

Except for clients, all OWS architecture components provide services to clients and/or to other components. Each such component is usually called a service when multiple implementations are expected, and each implementation is called a server (or service instance). These components are thus usually called services or servers in this chapter.

Clients are software packages that provide access to a human user or operate as agents on behalf of other software. Software that provides access to a human user can range from a web browser to a monolithic application with specific tailoring to the users needs.

All services (or components) are loosely organized in four tiers.

- Client tier
- Application Services tier
- Processing Services tier
- Information Management Services tier

This organization is loose in that clients and services can bypass un-needed tiers. Services can use other services within the same tier, and this is common especially in the Processing Services tier. (This is further described in the previous section OWS Fundamentals section 1b and 1c.) Also, some services perform functions of more than one tier, when those functions are often used together and combined implementation is more efficient. Assignment of such combined services to tiers is somewhat arbitrary.

This OWS architecture is designed for use where data is important and often voluminous. Servers can operate on (tightly bound) data stored in that server and/or on (loosely bound) data retrieved from another server. Most data is stored by the servers in the Information Management Services tier, but some data (can be and often) is stored in other services and servers.

Application Services Tier

The Application Services tier contains services designed to support Clients, especially thin client software such as web browsers. That is, these Application Services are designed for use by clients instead of each client directly performing these often-needed support functions. The services in the Application Services tier are used by Clients, and can use other services in the Application Services, Processing Services, and Information Management Services tiers. The specific services included in this tier include (but are not limited to) the services listed in Table 4.1.

Table 4.1 Some specific Application Services

Service name	Service description
Web portal services	Services that allow a user to interact with multiple application services for different data types and purposes
WMS application services	Services that allow a user to interact with a Web Map Service (WMS) to find, style, and get data of interest
Geographic data extraction services	Services that allow a user to extract and edit feature data, interacting with images and feature data
Geographic data management services	Services that allow a user to manage geospatial data input and retirement, interacting with Information Management Services
Chain definition services	Services to define a service chain and enable it to be executed by the workflow enactment service; may also provide a chain validation service
Workflow enactment services	Services to interpret chain definitions and control instantiation of servers and sequencing of activities, maintaining internal state information associated with various services being executed

Processing Services Tier

The Processing Services tier contains services designed to process data, sometimes both feature and image (coverage) data. The services in the Processing Services tier are used by clients and by services in the Application Services tier. These services can use other services in the Processing Services and Information Management Services tiers. The specific services included in this tier include (but are not limited to) the services listed in Table 4.2.

Information Management Services Tier

The Information Management Services tier contains services designed to store and provide access to data, normally handling multiple separate datasets. In addition,

Table 4.2 Some specific Processing Services

Service name	Service description
Web Coordinate Transformation Service (WCTS)	Transforms the coordinates of feature or coverage data from one coordinate reference system (CRS) to another, including "transformations", "conversions", rectification, and orthorectification
Web Image Classification Service (WICS)	Performs classification of digital images, using client-selected supervised or unsupervised image classification method
Feature Portrayal Service (FPS)	Dynamically produces client-specified pictorial renderings in an image or graphics format of features and feature collections usually dynamically retrieved from a Web Feature Server (WFS)
Coverage Portrayal Service (CPS)	Dynamically produces client-specified pictorial renderings in an image or graphics format of a coverage subset dynamically retrieved from a Web Coverage Service (WCS)
Geoparser Service	Scans text documents for location-based references, such as a place names, addresses, postal codes, etc., for passage to a geocoding service.
Geocoder Service	Service to augment location-based text references with position coordinates
Dimension measurement services	Services that compute dimensions of objects visible in an image or other geospatial data
Route determination services	Determine optimal path between two specified points based on input parameters and properties contained in a Feature Collection; may also determine distance between points and/or time to follow path
Change detection services	Services to find differences between two data sets that represent the same geographical area at different times
Feature generalization services	Service that reduces spatial variation in a feature collection to counteract the undesirable effects of scale reduction
Format conversion services	Converts data from one format to another, including data compression and decompression

Table 4.3 Some specific Information Management Services

Service name	Service description
Web Map Service (WMS)	Dynamically produces spatially referenced maps of client-specified ground rectangle from one or more client-selected geographic datasets, returning pre-defined pictorial renderings of maps in an image or graphics format
Web Feature Service (WFS)	Retrieves features and feature collections stored that meet client-specified selection criteria
Web Coverage Service (WCS)	Retrieves client-specified subset of client-specified coverage (or image) dataset
Catalog Service for the Web (CSW)	Retrieves object metadata stored that meets client-specified query criteria
Order handling services	Allows clients to order products from a provider, including: selection of geographic processing options, obtaining quotes on orders, submission of order, statusing of orders, billing, and accounting

metadata describing multiple datasets can be stored and searched. Access is usually to retrieve a client-specified subset of a stored dataset, or to retrieve selected metadata for all datasets whose metadata meets client-specified query constraints.

The services in the Information Management Services tier are used by clients and by services in the Application Services and Processing Services tiers. These services can use other services in the Information Management Services tier. The specific services included in this tier include (but are not limited to) the services listed in Table 4.3.

4.2.1.3 Service Trading (Publish – Find – Bind)

All OGC architecture services are self-describing, supporting dynamic (just-in-time) connection binding of servers using service trading. Service trading addresses discovery of available service instances. Trading facilitates the offering and the discovery of interfaces that provide services of particular types. A trader implementation records service offers and matches requests for advertised services. Publishing a capability or offering a service is called “export”. Matching a service request against published offers or discovering services is called “import”. This can also be depicted in an equivalent manner as the “Publish – Find – Bind” (PFB) pattern of service interaction. The fundamental roles are:

1. Trader (Registry) – registers service offers from exporter objects and returns service offers to importer objects upon request according to some criteria.
2. Exporter (Service) – registers service offers with the trader object
3. Importer (Client) – obtains service offers, satisfying some criteria, from the trader object.

In the OWS architecture, a Registry is implemented using the OpenGIS® Catalog Service for the Web (CSW) Interface Standard.

A trader plays the role of "matchmaker" in a service-oriented architecture. The interaction pattern is:

- To publish a service offer, an Exporter gives a Trader a description of a server, including a description of the interface at which that service instance is available.
- To find suitable server offers, an Importer asks a Trader for a server having certain characteristics. The trader checks the previously registered descriptions of servers, and responds to the importer with the information required to bind with a server. Preferences may be applied to the set of offers matched according to service type, constraint expressions, and various policies. Use of preferences can determine the order used to return matched offers to the importer.
- To bind a service, an Importer applies information received from the Trader to bind to a server. The Client then proceeds to use that server.

4.2.1.4 SOAP and REST

OGC anticipates ongoing changes and evolution in the distributed computing platforms (DCPs) in which the OGC standards are based. Geospatial operations and information concepts are not directly affected by the change in development DCPs but the implementation specifications must be written for specific platforms. OGC's strategy is that the abstract models for geospatial services are mapped onto the various DCPs.

Currently for web services, the most important DCP discussions involve SOAP and REST, which define different approaches to implementing services in a web environment. SOAP, originally defined as Simple Object Access Protocol, is a protocol specification for exchanging structured information in the implementation of web services. Representational state transfer (REST) is a style of software architecture for distributed hypermedia systems such as the World Wide Web. Development and initial adoption of OGC's web services standards predated the development of SOAP and REST approaches.

In July 2006, OGC members agreed on a strategy that future revisions of existing and new OWS interface specifications must include an optional SOAP binding. Several OGC Interoperability Program initiatives developed approaches to using SOAP with OWS specifications. Recently SOAP profiles have been added to several OGC standards.

At the July 2007 OGC meetings in Paris, the members agreed to form a REST Subcommittee that would work on developing best practices guidance related to the use of OGC web services in a RESTful environment.

The consensus is that there is not an either/or decision related to "REST vs. SOAP". Instead, we need best practices and guidance for both architecture patterns. There is agreement that both REST and SOAP have their strengths and weaknesses and that the real question is when to use either approach – or at times a blended approach.

4.2.1.5 Service Chaining

In many cases, multiple services must be used together to perform a useful function. The OWS architecture thus supports "chaining" together of multiple servers, and such chaining is frequently used. This chaining is not limited to a linear chain; a network of services can also be "chained". Within such a chain, most servers input the data that is output from the previous server in the chain. Services can be chained transparently (defined and controlled by the client), translucently (predefined but visible to the client), and opaquely (predefined and not visible to client), see Subclause 7.3.5 of [OGC 02-006, ISO 19119].

To facilitate service chaining, some services support defining and executing chains of services. Also, some Processing Service interfaces are designed to support retrieving the data to be processed from another service, which can be an Information Management Service or another Processing Service.

To allow more efficient execution of server chains, some service interfaces support server storage of operation results until requested by next service in a chain. This approach separates the flow of control from the flow of data.

4.2.1.6 Service Communication

Communication between clients and services, and between services, uses only open non-proprietary Internet standards. That is, the OWS architecture uses the Internet or equivalent as its distributed computing platform (DCP). More specifically, communication between components uses standard World Wide Web (WWW) protocols, namely HTTP GET, HTTP POST, and SOAP. Specific operations of specific servers are addressed using Uniform Resource Locators (URLs). Multipurpose Internet Mail Extensions (MIME) types are used to identify data transfer formats. The data transferred is often encoded using the Extensible Markup Language (XML), with the contents and format carefully specified using XML Schemas.

4.2.1.7 Service Interfaces

OGC web service interfaces use open standards and are relatively simple. All services support open standard interfaces from their clients, often OGC-specified service interfaces. In addition to being well-specified and interoperable tested, the OGC-specified service interfaces are coarse-grained, providing only a few static operations per service. For many services, only three service operations are specified. One server can implement multiple service interfaces whenever useful.

The OGC web service interfaces are usually stateless, so session information is not passed between a client and server. Clients retain any needed interface state between operations.

The OGC web service interfaces share common parts whenever practical, allowing those parts to be specified and implemented only once. For example, all OWSs have a mandatory GetCapabilities operation to retrieve server metadata. That server

Table 4.4 Some standardized encoding formats and languages

Specification name	Description
Styled Layer Descriptor (SLD)	Encodes client-controlled styling for map portrayal of features and coverages (images)
Geography Markup Language (GML)	Language defined using XML Schemas based on the ISO 191XX series of standards, to be used to specify application-specific XML Schemas
Coordinate Reference Systems (part of GML)	Encodes definitions of coordinate reference systems, coordinate systems, datums, and coordinate transformations (and conversions)
OWS Context	Encodes multiple OWS application display context
URNs using ogc URN namespace	Standardized Universal Resource Identifiers (URNs) referencing most well-known coordinate reference systems (CRSs) and grid CRSs
Web Service Description Language (WSDL)	Encodes web service interfaces
Business Process Execution Language (BPEL)	Encodes sequences of interactions with seb services. BPEL scripts are developed and executed to produce information for specific purposes

metadata includes four required sections, with the contents and format of three sections common to all services, and part of the fourth section common to most services. In addition, many service interfaces have multiple specified levels of functional compliance, or multiple specialized subset and/or superset profiles.

Standard XML-based data encoding formats and languages are used in many server-to-client and client-to-server data transfers. The formats and languages specified include (but are not limited to) those listed in Table 4.4. In these formats and languages and elsewhere, the geographic data and service concepts are closely based on the ISO 191XX series of standards.

4.2.1.8 Server Implementation

Servers and client implementations are not constrained except for supporting the specified service interfaces. Each can be implemented by software executing on any general-purpose computer connected to the Internet or equivalent. The architecture is hardware and software vendor neutral. The same and cooperating services can be implemented by servers that are owned and operated by independent organizations.

All OWS services and clients are implemented by available standards-based Commercial Off The Shelf (COTS) software. This commercial software can sometimes be used without requiring major software development, or can be adapted to specific needs with limited software development. Software may be developed as proprietary or open source code.

4.3 OWS Testbeds

Earth observation interoperability requirements are a critical underpinning of OWS standards. These standards are usually developed and refined in the OGC Interoperability Program's testbeds, pilot projects, and interoperability experiments. Sponsoring organizations fund these activities in which participating organizations develop specifications and software components, test interoperability with other components, and produce documentation. Documents include Discussion Papers, Engineering Reports, Best Practices, Specification Profiles, Change Requests, and Reference Models as well as draft standards. These are reviewed and approved (approval is often contingent upon submitters completing specific improvements) by the Technical Committee and Management Committee.

Most new OWS standards have come from the OWS testbeds – OWS-1, OWS-2, OWS-3, OWS-4, OWS-5, and OWS-6. Other testbeds have been sponsored by organizations or teams of organizations with specific interoperability needs or particular needs for a separate initiative. (A full list of current and past initiatives is available online http://www.opengeospatial.org/projects.)

In a testbed, the design, development and testing of components and specifications is typically conducted over a 6-month period preceded by a call for sponsors, a request for quotations and a participant selection process. A testbed usually concludes with a demonstration of interoperability involving a variety of commercial and prototype products and components in a realistic scenario. Most of the demonstrations have been captured in multimedia, and videos are available on the OGC website. OWS-1 and OWS-2 results are available on the website.

4.3.1 OWS-3

The OWS-3 initiative, in 2006, was organized around the following threads:

1. Common Architecture: The Common Architecture (CA) thread addressed issues, infrastructure and requirements necessary to integrate services implemented using OGC specifications into an operational Web Service enterprise. For OWS-3, the emphasis of the Common Architecture thread was on capturing best practices, extending the scope and capabilities of catalog services, and maturing OWS workflow.
2. Sensor Web Enablement (SWE): The Sensor Web Enablement (SWE) thread matured the existing set of SWE work items to enable the federation of sensors, platforms and management infrastructure into a coordinated sensor enterprise. This enterprise will enable the discovery and tasking of sensors as well as the delivery of sensor measurements regardless of sensor type and controlling organization.
3. Geo-Decision Support Services (GeoDSS): Geo-Decision Support Services (GeoDSS) built on the Information Interoperability work from OWS-2 to explore

ways to tailor geographic information for different information communities. GeoDSS refined and extended the OGC Portrayal encoding and services through application to symbology encodings from two communities. In addition, GeoDSS developed the new capability of a Geo-Video Service (GVS). Finally, GeoDSS explored extensions/enhancements to the underlying OGC services to address a greater extent of emergency response scenarios.
4. Open Location Services (OpenLS): OpenGIS Location Services (OpenLS) comprise an open platform for position determination and location-based applications targeting mobile terminals.
5. Geo-Digital Rights Management (GeoDRM): The Geospatial Digital Rights Management (GeoDRM) thread in OWS-3 was the first step of adding digital rights protocols to the existing OWS architecture. The GeoDRM thread in OWS-3 extended the "click-through" licensing concept for web sites to geospatial data services. In particular, click-through licensing techniques were developed for the Web Map Service and Web Feature Service.

4.3.2 OWS-4

The OGC Web Services, Phase 4 (OWS-4) Testbed (June to December 2006) had 11 sponsoring organizations who responded to a January 2006 Call for Sponsors and defined a set of requirements. Seventy two (72) organizations who responded to an April 2006 Request for Quotations (RFQ) and Call for Participation (CFP) participated in some aspect of OWS-4. Fifty nine (59) components were implemented and deployed in interoperability testing in seven threads:

1. Sensor Web Enablement (SWE): The implementation and testing of SWE components reached a level of maturity sufficient to support the adoption of SWE specifications as standards by the OGC Technical Committee at the level of Version 1.0, i.e., O&M, TML, SensorML, SOS, SPS.
2. Geo Processing Workflow (GPW): A baseline approach for OWS Workflow using BPEL was established and demonstrated in several scenarios. Several processing services were defined as profiles of the Web Processing Service, e.g., Topology Quality Assessment Service, Model Output Processing Service.
3. Geo-Decision Support (GeoDSS): An open-source GML Client Application was developed and released as part of the OWS-4 DVD and through Source Forge. While this application is limited in GIS functionality it provides geospatial browsing, supporting the visual integration of GML with WMS and WFS services. Guidance for mapping domain models to the ebRIM (electronic business Registry Information Model) model for CSW was developed. Progress was made on techniques for developing GML Application Schemas. The OWS approach to Portrayal was refined including separation of the SLD specification into two parts: Symbology Encoding (SE) and SLD profile of WMS. Also clearly identified were the two services of Feature Portrayal Service (FPS) and

the Integrated SLD-WMS. The WCS was extended to accommodate a response using JPIP (JPEG 2000 Interactive Protocol) image streaming, i.e., geo-enabling JPIP. The WCS parameters for requesting geospatial coverages provided an enhancement to the efficiency of JPIP data transfer.

4. An initial architecture profile of OGC Web Services for the National System for Geospatial-Intelligence (NSG) was developed.
5. Geo-Digital Rights Management (GeoDRM): Implementation of GeoDRM and Security components consistent with the OGC GeoDRM Abstract Specification was accomplished. The implemented architecture was captured in an Engineering Viewpoint architecture.
6. CAD / GIS / BIM (CGB): OWS-4 was the first web services implementation of a set of CAD-GIS-BIM requirements; initial discovery and access to CGB data was achieved by extending existing OWS specifications. An architecture for further development was defined.
7. OGC Location Services (OpenLS)
8. Compliance Testing (CITE): Compliance Test Scripts and Reference Implementations for SDI 1.0 were developed, as well as a new open source test engine (the TEAM engine)

At the OWS-4 demonstration, held at an Emergency Operations Center (EOC) in the New York/New Jersey metropolitan area, high level disaster managers from state, federal and local agencies saw live Web-based information systems being used to find, access and integrate diverse geospatial resources, just as these managers' systems might be used in a real disaster. The information flowed from many different data sources, most using commercially available off-the-shelf software implementing OGC standards. In the demo, the following capabilities were shown:

- The OGC's SWE Standards made it possible to find and control online sensors as diverse as radiation counters, anemometers, security cameras, and NASA imaging satellites. An operator in the demo accessed NASA's Earth Observation-1 (EO-1) satellite ground system, instructing the satellite through an open interface to provide images of the New York/New Jersey area over the next several days. The acquisition request was accepted by the EO-1 planning systems and the image was acquired on December 8th during the OWS-4 demonstration. NASA satellites are in fact being fitted with the open SWE specifications to make such use possible.
- Commercial weather data sources and weather forecasts were also accessed. Using information from these and from wind sensors, a radioactivity dispersion plume was calculated, and within less than 1 h managers at the EOC had begun a fictional evacuation of areas that had been or would be impacted.
- Service chaining for decision support made it possible to, in effect, create "macros" or packaged sets of services hosted on multiple remote servers, in order to streamline the delivery of information to decision makers.

- Proposed geospatial digital rights management (GeoDRM) standards enabled emergency access to sources of data that were either private and proprietary or public but under legal constraints.
- Geosemantics applied to metadata in catalogs played a role in discovering the best available data and services.
- Multi-lingual data and map services allowed participants with different native languages to collaborate more effectively.
- Building information models (BIM) integrated computer-aided design (CAD) data with geospatial data and text made it possible to review and compare different buildings to choose the one most suitable for an emergency field hospital.

4.3.3 OWS-5

OGC Web Services, Phase 5 (OWS-5) Testbed (July 2007 to April 2008): 7 Sponsoring organizations who responded to a Call for Sponsors defined the requirements and developed the scenario for OWS-5. Thirty five 35 organizations who responded to a May 2007 Request for Quotations (RFQ) and Call for Participation (CFP) participated. Fifty two (52) components were implemented and deployed in five threads:

1. Sensor Web Enablement (SWE): Participants demonstrated implementation and integration of IEEE-1451 TIM (Transducer Interface Module), NCAP (Network Capable Application Processor) and STWS (Smart Transducer Web Services) components and refined the integration of IEEE-1451 sensors into the SOS framework. A BPEL script was developed for SWE GeoReferenceable workflow. This workflow established a standardized means to allow the user to interactively access a subset of pixels from a coverage service stored in the compressed JPEG2000 and preserve the image relationship with the associated "sensor" model parameters such that precise geopositioning capabilities could be realized in a dynamic, interactive and networked environment. The OGC specifications used in this scenario included: JPIP enabled WCS-T 1.1, CS/W, WPS, SPS, SAS, and SOS.
2. Geo Processing Workflow (GPW): Participants developed SOAP and WSDL interfaces for four foundation services: WMS, WFS-T, WCS-T, and WPS, allowing these services to be integrated into industry standard service chaining tools. Service Implementations for WFS-T, WCS-T, WMS and WPS were deployed to demonstrate SOAP and WSDL binding patterns. They developed a BPEL script for SWE GeoReferenceable workflow. This workflow establishes a standardized means to allow the user to interactively access a subset pixels from a coverage service stored in the compressed JPEG2000 and preserve the image relationship with the associated sensor model parameters such that precise geopositioning capabilities can be realized in a dynamic, interactive and networked environment.

The OGC specifications used in this scenario included: JPIP-enabled WCS-T 1.1, CS/W, WPS, SPS, SAS, and SOS.

3. Geo-Decision Support (GeoDSS): Participants demonstrated feasibility of the draft Web Coverage Processing Service (WCPS) standard by implementing use cases (sensor time-series, oceanography, remote sensing imagery.) A Conflation workflow process and BPEL script were designed and implemented to demonstrate service chaining and workflow, web processing services, and service interoperability using a variety of OGC service standards. There was successful design, implementation. and testing of data view models harvested in a catalog. The UML (Unified Modeling Language)-GML Application Schema (UGAS) tool was enhanced to include: utilization of OCL constraints; schema generation based on ISO/TS 19139 encoding rules; and capability to integrate existing XML grammars based on XML attributes.
4. Agile Geography: Participants specified how KML could be output from a geospatial database using three existing standards: WMS for the overall information request, WFS Filter for the query, and SLD for styling rules. They completed a proposal for a new OGC standard for Federated Geo-synchronization and developed an abstract core WFS module and a series of other modules that instantiate Web-based data provisioning.
5. Compliance Testing (CITE): The Compliance and Interoperability Test and Evaluation (CITE) thread developed 6 compliance test suites

4.3.4 OWS-6

The OGC Web Services, Phase 6 (OWS-6) Testbed Call for Sponsors took place in July 2008. The testbed was just getting underway at the time of this writing and was due to complete in April 2009. The five planned threads are:

1. Sensor Web Enablement (SWE): OWS-6 will focus on integrating the SWE interfaces and encodings into cross-thread scenarios and workflows to demonstrate the ability of SWE specifications to support operational needs. Emphasis will be on:

   - Integrating CCSI-Enabled CBRN Sensors into the SWE Environment
   - Harmonizing SWE-related information models: SensorML, GML, UncertML, MathML
   - Applying GeoRM, Trusted Services, and security models in SWE environment
   - Events-based architecture including WNS

2. Geo Processing Workflow (GPW): This GPW thread aims to build on the progress of previous testbeds with particular emphasis on service security issues. To satisfy mission-critical goals, the architecture must ensure authenticity,

integrity, quality and confidentiality of services and information. The following task areas have been identified:

- Asynchronous Workflow and Web Services Security
- Data Security for OGC web services
- Data Accessibility
- WPS Profiles - Conflation; and Grid processing
- GML Application Schema Development and ShapeChange Enhancements

3. Aeronautical Information Management (AIM): The Aviation Information Management (AIM) subtask is a new thread within OWS to develop and demonstrate the use of the Aeronautical Information Exchange Model (AIXM) in an OGC Web Services environment. This thread will focus on evaluating and advancing AIXM features in a realistic trans-Atlantic aviation scenario, devising and prototyping a Web Services Architecture for providing valuable aeronautical information directly to flight decks, Electronic Flight Bags (EFB) and handheld devices (such as PDAs and Blackberries) while the airplane is at the gate or en-route to its destination. AIXM was developed by the Federal Aviation Administration (FAA) and Eurocontrol as a global standard for the representation and exchange of aeronautical information, enabling the transition to a net-centric, global aeronautical management capability. It uses the ISO 19100 modeling framework and has two major components: a conceptual model presented in the form of an UML class model and a data encoding specification which was developed using the OGC Geography Markup Language (GML). Both have been tailored for the representation of aeronautical objects, especially the temporality feature that allows for time-dependent changes affecting AIXM features. The OWS-6 AIM thread shall perform tasks in the following areas:

   - Use and enhancement of Web Feature Service and Filter Encoding specifications in support of AIXM features and 4-dimensional flight trajectory queries,
   - Prototype of Aviation client for retrieval and seamless visualization of AIXM, Weather and other aviation-related data, emphasizing time and spatial filtering in order to present just the right information into a given user context anytime, anywhere,
   - Architecture of the standards-based mechanism to notify users of changes to user-selected aeronautical information.

4. Decision Support Services (DSS): Decision Support Services involving geospatial and temporal information has been a recurring thread in OWS testbeds. This thread focuses on presenting and interacting with data obtained from the sensor web and geoprocessing workflows to support analysis and decision making. The focus for DSS in OWS-6 builds on portrayal, WMS Tiling, and integrated client work from OWS-4, with additional work on 3D visualization and integration of the built environment and landscape. This thread will encompass:

   - ISO 19117 and OGC SLD Portrayal
   - 3D Portrayal of GML with Fly-through

- Hosting CityGML data with WFS
- Outdoor and indoor 3D route and tracking services
- WMS performance (tiling)
- Integrated Client for multiple OWS services

5. Compliance and Interoperability Test and Evaluation (CITE): The major geospatial industry consumers require verifiable proof of compliance with OGC specifications in order to reach the desirable outcome of interoperability. The CITE threads in previous OWS projects have made significant progress towards having a complete suite of compliance tests for this baseline of interfaces. A major focus of OWS-6 CITE will be in clearly documenting the approach to defining Abstract Test Suites. In addition, OWS-6 will expand the usability of the existing OGC compliance tests by "tailoring" these tests for specific schema profiles and/or data.

4.4 Conclusion

Existing OGC specifications are widely implemented in Earth observation software and the larger geospatial technology marketplace. The specifications for the Information Management Tier (WMS, WFS, WCS, CSW) are mature with multiple implementations. Emphasis in the OGC development activities is focused on developing Processing Services and best practices for service chaining, e.g., workflow. The OWS architecture is also the basis for the recently approved OGC Sensor Web Enablement (SWE) set of standards for accessing any type of sensor as a web service. Approaches for managing digital rights in the OWS environment are being developed, which will help overcome many of the institutional and commercial obstacles to data sharing and market growth. Concepts are under development for applying the OWS architecture and services in a mass market environment, which puts technically sophisticated services in the hands of non-technical users and opens up a much larger market for geospatial technology and data providers. This progress depends on cooperation among technology users and competing technology providers, and OGC's function is to foster and manage such cooperation.

Chapter 5
Evolution of the Earth Observing System (EOS) Data and Information System (EOSDIS)

Hampapuram K. Ramapriyan, Jeanne Behnke, Edwin Sofinowski, Dawn Lowe, and Mary Ann Esfandiari

5.1 Introduction

One of the strategic goals of the US National Aeronautics and Space Administration (NASA) is to "Develop a balanced overall program of science, exploration, and aeronautics consistent with the redirection of the human spaceflight program to focus on exploration" (NASA 2006). An important sub-goal of this goal is to "Study Earth from space to advance scientific understanding and meet societal needs." NASA meets this sub-goal in partnership with other US agencies and international organizations through its Earth science program. A major component of NASA's Earth science program is the Earth Observing System (EOS). The EOS program was started in 1990 with the primary purpose of modeling global climate change. This program consists of a set of space-borne instruments, science teams, and a data system. The instruments are designed to obtain highly accurate, frequent, and global measurements of geophysical properties of land, oceans, and atmosphere. The science teams are responsible for designing the instruments as well as scientific algorithms to derive information from the instrument measurements. The data system, called the EOS Data and Information System (EOSDIS), produces data products using those algorithms and then archives and distributes such products. The first of the EOS instruments were launched in November 1997 on the Japanese satellite called the Tropical Rainfall Measuring Mission (TRMM) and the last, on the US satellite Aura, were launched in July 2004. The instrument science teams

H.K. Ramapriyan (✉)
NASA Goddard Space Flight Center, Greenbelt, MD, USA
e-mail: rama.ramapriyan@nasa.gov

This work was performed by the first two and the fourth and fifth authors as part of their official duties as employees of the US government. It was supported by the NASA's Science Mission Directorate. The third author worked as a contractor supporting this effort under contract NNG05CA99C between NASA and SGT, Inc. The opinions expressed are those of the authors and do not necessarily reflect the official position of NASA.

L. Di, H.K. Ramapriyan (eds.), *Standard-Based Data and Information Systems for Earth Observation*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-540-88264-0_5,

have been active since the inception of the program and have participation from Brazil, Canada, France, Japan, Netherlands, the United Kingdom, and the US. The development of EOSDIS was initiated in 1990, and this data system has been serving the user community since 1994. The purpose of this chapter is to discuss the history and evolution of EOSDIS since its beginnings to the present and indicate how it continues to evolve into the future. See Ramapriyan (2003) for a more detailed discussion of the history.

In the 1980s, NASA's Earth science data were generally held by principal investigators or held at specialized data systems. Access to data and data products was limited to the individual scientist or small team responsible for generating the data. There were no policy-driven requirements for principal investigators to make their data available to other scientists or to a broader user community until the end of their missions. For the Upper Atmospheric Research Satellite (UARS) mission, NASA established a more open data policy whereby two years after the start of the mission the data were publicly available.

An even more open data policy was adopted by NASA for EOS. According to the EOS data policy, whose goal was to make the data available to a broad community, there was to be no exclusive access to data after an initial checkout period (EOS Project Science Office 1990). A set of "standard products" was defined for each of the instruments on the EOS spacecraft. The EOS instrument teams would develop these products using peer-reviewed algorithms and make them available to all users for research and applications. Recognizing the importance of data management, NASA started the Earth Science Data and Information System (ESDIS) Project separately from the projects responsible for the spacecraft and instruments. The purpose of the ESDIS Project was to develop and operate EOSDIS. Initially, there were to be two sites constituting EOSDIS – NASA's Goddard Space Flight Center (GSFC) and the United States Geological Survey's (USGS) Earth Resources Observation Systems (EROS) Data Center (EDC) for data processing, archiving, and distribution. However, given the variety of Earth science disciplines covered by the EOS Program, it was not practical for two data centers to have the necessary expertise and understanding of the data needed to serve the scientific community effectively. This called for a larger number of data centers, distributed throughout the country to take advantage of existing scientific and data management expertise (Science Advisory Panel 1990).

In 1990, NASA selected several organizations in the US based on scientific disciplines and heritage data management expertise. These were named Distributed Active Archive Centers (DAACs) since these data centers would provide a stable repository or archive of the EOS data products, actively manage the data in the repository, and would be distributed across the country. As the DAACs were established, it was also recognized that making heritage data more easily available to the community would be good preparation for managing the large data flows from EOS. The initial version of EOSDIS to accomplish this was called Version 0 (V0), a "working prototype with operating elements" (Ramapriyan and McConaughy 1991). Developed collaboratively by the DAACs and the ESDIS Project to improve access to existing data at the DAACs, V0 was operationally released to users

in August 1994 and adopted a World Wide Web (WWW) interface within three months thereafter. Also, tailored interfaces were developed by DAACs to serve their individual discipline communities.

In parallel with the V0 development, the ESDIS Project was preparing to satisfy the requirements for "big" data flows from the EOS missions through two major subsystems. The EOS Data and Operations System (EDOS) would be developed for data capture and initial (Level 0) processing. The EOSDIS Core System (ECS) would satisfy the remaining functions of flight operations (command and control of spacecraft and instruments) through its Flight Operations Segment (FOS) and the processing, archiving and distribution of the data from the EOS instruments through its Science Data Processing Segment (SDPS). The SDPS would perform all the functions past Level 0 processing of data from all EOS instruments (starting with those on the Tropical Rainfall Measuring Mission – TRMM – scheduled for launch in 1996) and would also support all the heritage data that were being managed using V0. As development of SDPS progressed, it became clear that the system was too complex with too many requirements. Over the period 1995 through 1999, actions were taken to decentralize the development and simplify the system in order to meet the objectives of the EOS mission. Systems based on V0 at the DAACs were used to support TRMM. Generation of standard products from most of the instruments' data was moved to Science Investigator-led Processing Systems (SIPSs) that would be developed and operated by the respective instrument teams. An EOS Data Gateway (EDG) would be developed based on V0 IMS. The remaining functions in the SDPS were prioritized with input from the scientific user community, and releases of SDPS were scheduled to occur frequently with demonstrably increasing functionality with each release. This led to the successful completion of all subsystems needed to support the Terra mission (launched in December 1999) on time.

Since 1999, EOSDIS has been supporting ingest, processing, archiving and distribution of all the data from the EOS instruments and the products derived from them. The missions and instruments supported by EOS are shown in Fig. 5.1. The original DAAC concept has progressed to include a variety of data centers. Today, EOSDIS Data Centers hold over 2700 distinct datasets that include EOS and heritage (pre-EOS) products. A large and diverse community has become accustomed to data and information products from EOSDIS, as evidenced by the number of users visiting EOSDIS web sites (over 450 thousand), receiving more than one and a half petabytes of data in 2007. At the end of 2007, EOSDIS archives held about 3.75 petabytes of data, growing at a rate of $\sim$1.7 terabytes per day.

An example of an EOS mission Earth science instrument is the Moderate Resolution Imaging Spectroradiometer (MODIS) flown on both the Terra and Aqua EOS mission satellites. This instrument provides data that improves our understanding of global dynamics and processes occurring on the land, in the oceans, and in the lower atmosphere. The two instances of this instrument contribute to a significant proportion of the EOSDIS data processing archive and distribution resources.

In 2005, after more than 10 years in operation, it was time to re-examine lessons learned and seek significant improvements in a variety of areas. NASA

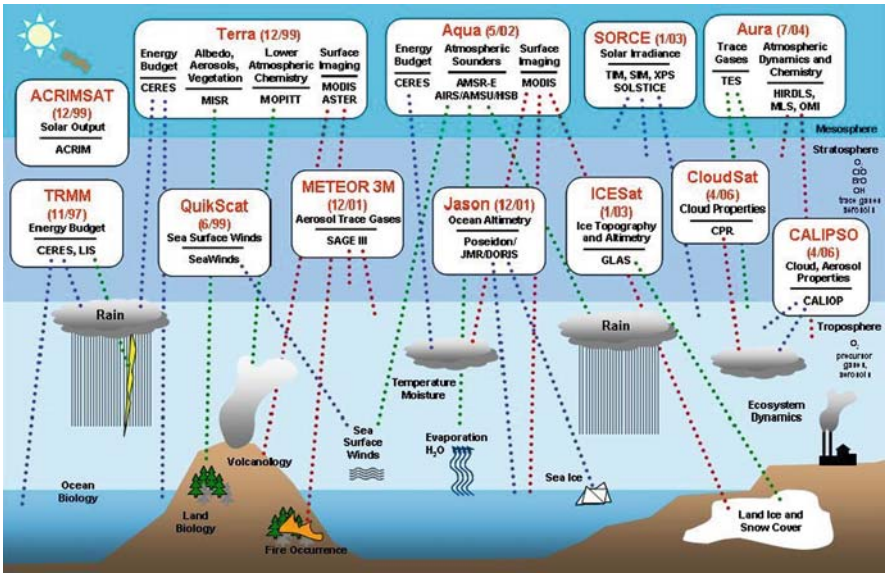

Fig. 5.1 Missions and instruments supported by EOSDIS

established an EOSDIS Evolution Study to develop an approach and implementation plan that would begin to fulfill the objectives set forth in a vision for circa 2015.

The remainder of this chapter is organized as follows. Section 5.2 provides a discussion of EOSDIS, its elements and their functions. Section 5.3 provides details regarding the move towards more distributed systems for supporting both the core and community needs to be served by NASA Earth science data systems. Section 5.4 discusses the use of standards and interfaces and their importance in EOSDIS. Section 5.5 provides details about the EOSDIS Evolution Study. Section 5.6 presents the implementation of the EOSDIS Evolution plan. Section 5.7 briefly outlines the progress that the implementation has made towards the 2015 Vision, followed by a summary in Sect. 5.8.

5.2 EOSDIS and Its Elements

The EOSDIS is a geographically distributed, end-to-end data system for command and control of EOS spacecraft and instruments; for receipt, capture, and Level 0 processing of telemetry data; and for production, archival, and distribution of science data. It includes the communications and administration infrastructure necessary to "glue" the system together and monitor its operation. The parts of the system that perform functions starting with command and control and ending in Level 0 processing constitute the EOSDIS Mission Systems. See Fig. 5.2. The remaining elements

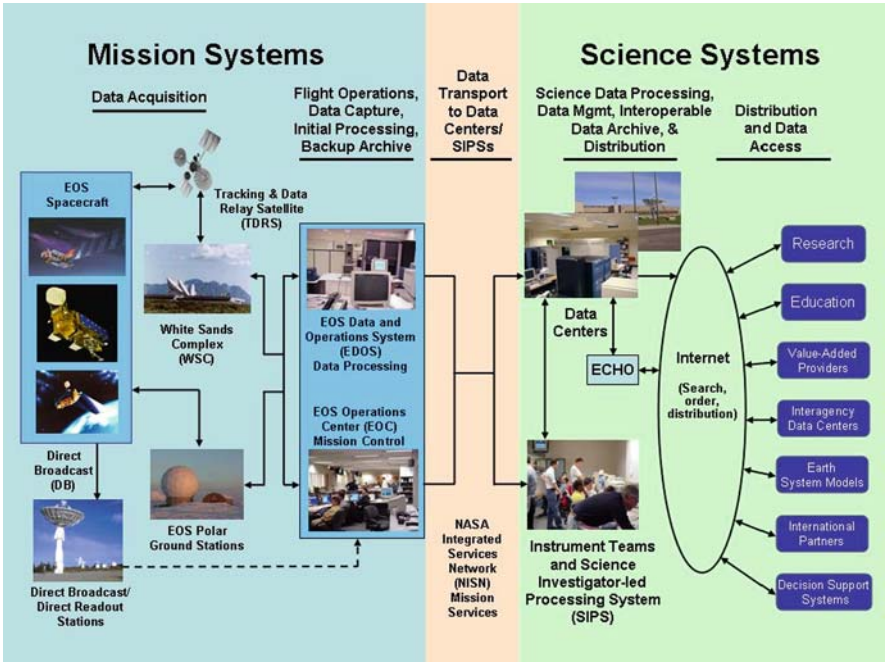

Fig. 5.2 EOSDIS missions and science systems

constitute the EOSDIS Science Systems maintained and operated by the NASA ESDIS Project. These science system elements were the focus of the EOSDIS Evolution study.

EOSDIS Mission Systems monitor the EOS spacecraft and instruments and ensure that the science data reach the ground systems. The mission system Level 0 production facility, the EOS Data and Operations System (EDOS), is the primary interface to the EOSDIS Science Systems. Level 0 data reaches the science systems through the NASA Integrated Services Network.

5.2.1 EOSDIS Data Centers

The twelve geographically distributed EOSDIS Data Centers – four at the Goddard Space Flight Center in Maryland and eight distributed around the rest of the United States – are collocated with other institutional facilities to achieve science synergy with the ongoing activities of those institutions. Each data center is responsible for EOSDIS data management and user services functions within a particular discipline area, as presented in Table 5.1. These data centers are located throughout the US (see Fig. 5.3):

Table 5.1 EOSDIS data centers

Data center	Location	Science disciplines
Alaska Satellite Facility (ASF) Distributed Active Archive Center (DAAC)	University of Alaska, Fairbanks, AK	Synthetic Aperture Radar (SAR) products, sea ice, polar processes, and geophysics
Crustal Dynamics Data and Information System (CDDIS)	NASA Goddard Space Flight Center, Greenbelt, MD	Space Geodesy and Geodetics
Global Hydrology Resource Center (GHRC)	NASA Marshall Space Flight Center, Huntsville, AL	Hydrologic cycle, severe weather interactions, lightning, and atmospheric convection
GSFC Earth Sciences (GES) Data and Information Services Center (DISC)	NASA Goddard Space Flight Center, Greenbelt, MD	Global precipitation, solar irradiance, atmospheric composition, atmospheric dynamics, global modeling
Land Processes (LP) DAAC	USGS EROS Data Center, Sioux Falls, SD	Land processes, land imaging
Langley Atmospheric Sciences Data Center (ASDC)	NASA Langley Research Center, Hampton, VA	Radiation budget, clouds, aerosols, and tropospheric chemistry
Level 1 and Atmospheres Archive and Distribution System (LAADS)/ MODIS Adaptive Processing System (MODAPS)	NASA Goddard Space Flight Center, Greenbelt, MD	MODIS level 1 and atmospheric data products
National Snow and Ice Data Center (NSIDC) DAAC	University of Colorado, Boulder, CO	Snow and ice, cryosphere, climate interactions and sea ice
Oak Ridge National Laboratory (ORNL) DAAC	Department of Energy Nashville, TN	Biogeochemical dynamics, ecological data, and environmental processes
Ocean Biology Processing Group (OBPG)	NASA Goddard Space Flight Center Greenbelt, MD	Ocean biology, sea surface temperature, and biogeochemistry
Physical Oceanography (PO) DAAC	Jet Propulsion Laboratory (JPL), Pasadena, CA	Sea surface temperature, ocean winds, circulation and currents and topography and gravity
Socioeconomic Data and Applications Center (SEDAC)	Columbia University, Palisades, NY	Human interactions, land use, environmental sustainability, geospatial data, multilateral environmental agreements

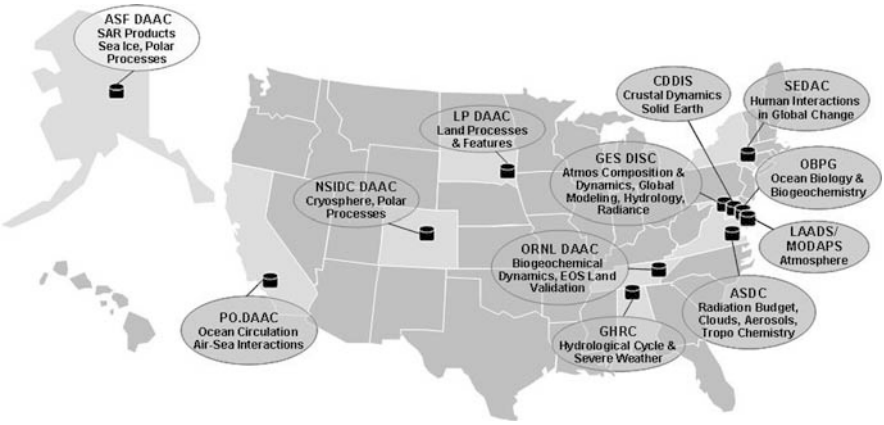

Fig. 5.3 EOSDIS data centers

The functions of the EOSDIS Data Centers are:

- Receiving EOS Level 0 data from the EOS Data and Operations System (EDOS)
- Receiving science software from EOS instrument teams and integrating it into an operational production environment
- Performing processing and reprocessing of standard data products following instrument teams' priorities
- Supporting science instrument teams as necessary in performing quality assurance of standard data products
- Ingesting standard data products produced at Science Investigator-led Processing Systems (SIPSs)
- Cataloging, archiving, and distributing EOS standard data products and other NASA Earth science data
- Providing data and information services and user support to the EOSDIS user community, and
- Preserving complete documentation of EOS data, instrument calibration, processing history, and processing source code

The EOSDIS Data Centers interface with other data centers and SIPSs to provide inputs for science product generation and to receive the science products for archiving and distribution. The EOSDIS Data Centers' primary purpose is to interact with science data users from around the world, providing access to NASA data. Each of the EOSDIS Data Centers has a Users Working Group (UWG) consisting of representatives of the user community in its particular scientific disciplines. The UWG provides the data center with advice on the priorities for the data sets and services offered by the data center.

5.2.2 Science Data Processing Segment

The Science Data Processing Segment (SDPS) performs information management and data archiving and distribution at each data center location. Each data center performs these functions using a combination of standard capabilities provided by the ESDIS Project and hardware and software specific to the data center. Special SDPS hardware and software, known as the EOSDIS Core System (ECS), was developed to support the high ingest rates of the EOS instruments. ECS currently resides and operates at three data centers: the Langley Atmospheric Science Data Center (ASDC), the Land Processes Distributed Active Archive Center (LP DAAC) and the National Snow and Ice Data Center (NSIDC). Data products are processed by the SIPSs or, in a few cases, by systems interfacing with the SDPS at the data centers. The SDPS at the data centers ingests the data from the processing systems and archives them. The SDPS has interfaces with the EOS Clearing House (ECHO) to provide search and access through ECHO clients, such as the Warehouse Inventory Search Tool (WIST). The SDPS also provides software toolkits to assist instrument teams in their development of product generation software at their Science Computing Facilities to facilitate ingest of the resulting products into SDPS or into data center-specific archiving and distribution systems.

5.2.3 Science Investigator-Led Processing Systems (SIPSs)

Most of the EOS standard products are produced at facilities under the direct control of the instrument Principal Investigators/Team Leaders (PIs/TLs) or their designees. These facilities are referred to as Science Investigator-led Processing Systems (SIPSs). The SIPSs are geographically distributed across the United States and are generally, but not necessarily, collocated with the PIs/TLs' Scientific Computing Facilities. Products produced at the SIPSs using investigator-provided systems and software are sent to appropriate EOSDIS Data Centers for archiving and distribution. Level 0 data products and ancillary data that begin the processing sequence are stored at the data centers and retrieved by the SIPSs. The geographic distribution of SIPSs is shown in Fig. 5.4.

5.2.4 EOS Clearing House (ECHO)

EOSDIS provides convenient mechanisms for locating and accessing products of interest. The "look and feel" of the system is intuitive and uniform across the multiple nodes from which EOSDIS can be accessed. EOSDIS facilitates collaborative science by providing extensible sets of tools and capabilities that allow investigators to provide access to special products (or research products) from their own computing facilities. The EOS Clearing House (ECHO) is a system developed by the ESDIS Project to provide a centralized spatial and temporal metadata collection of EOSDIS data. The fundamental principle of ECHO is to provide a central access path for any

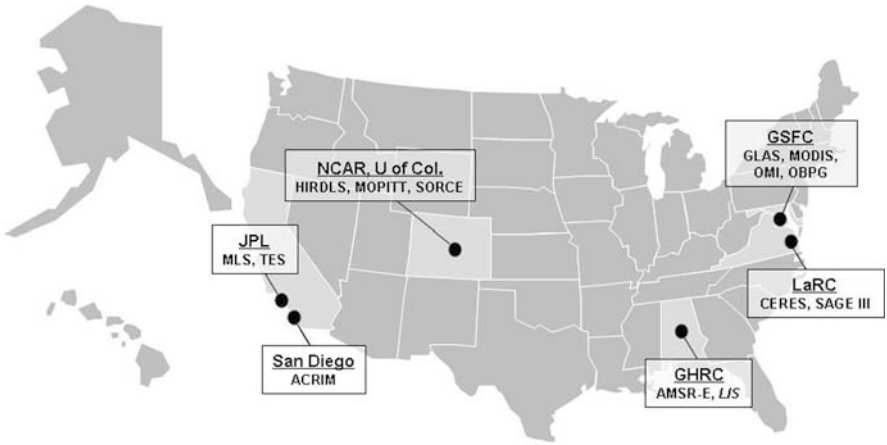

Fig. 5.4 EOSDIS science investigator-led processing systems

user interface developer, whether a NASA data system or any organization outside of NASA.

ECHO is the EOSDIS metadata framework by which EOSDIS keeps track of its vast data collection. ECHO is the middleware between EOS data and science data users via a service-oriented architecture. Data Partners provide metadata for their EOS data holdings and other Earth science-related data holdings. Client Partners develop software ("clients") to give science data users access to ECHO's registries using ECHO's open Application Programming Interfaces (APIs). Science data users search ECHO's registries and access data and services using an ECHO client. All of the EOSDIS Data Centers participate in ECHO by providing metadata to the ECHO database. One of the first user interfaces to be developed using ECHO is the Warehouse Inventory Search Tool (WIST), which provides web-based "one-stop shopping" for search and order capabilities within all of ECHO's data holdings. For more details about ECHO, see the ECHO web site (ECHO 2008).

5.3 Community Push Towards Distributed Systems

Since the early years of EOSDIS, there has been a push in the scientific community, with members from within and outside NASA, for a distributed implementation – from the points of view of both geography and responsibility. Two major influences from the community have introduced change in NASA's Earth Science data systems, including EOSDIS, over the last decade. These are recommendations from the National Research Council (NRC 1995), and the New Data and Information Systems and Services (NewDISS) Strategy Team (Maiden and NewDISS Team 2000).

In the mid-1990s, there was significant community concern about the centralized nature of the development of EOSDIS and doubts about its being able to meet all the requirements to satisfy the broader user community beyond the scientific researchers. The National Research Council reviewed EOSDIS in 1995 and recommended that the science data processing, archiving, and distribution should be performed by a "federation of competitively selected Earth Science Information Partners (ESIPs)" (NRC 1995). In response to this recommendation, NASA initiated an experiment in 1998 with a "self-governing" federation consisting initially of 24 competitively selected ESIPs, one-half (called Type 2 ESIPs) responsible for specialized research products and the other half (called Type 3 ESIPs) for products suitable for applications with commercial potential. The DAACs, whose primary responsibility was schedule-driven operational production and support of large user communities, were later included in the federation as Type 1 ESIPs. Initially sponsored by NASA, the ESIP Federation now consists of more than 110 members including NASA and NOAA data centers, research universities and laboratories, educators, technology developers and commercial and non-profit organizations. The Foundation for Earth Science was established in 2001 as a coordinating organization that promotes the objectives of the ESIP Federation, namely, bringing the most current and reliable data products based on satellite data to a broad range of users and ensuring their utilization to address environmental, economic, and social challenges of the world. Details about the ESIP Federation can be found on their web site (ESIPFED 2006).

NASA also commissioned, in 1998, the NewDISS Strategy Team with the charter to "define the future direction, framework, and strategy of NASA's Earth Science Enterprise (ESE) data and information processing, near-term archiving, and distribution." This team made a number of recommendations on how to proceed with ESE data and information systems and services over 6–10 years beyond the year 2000 (Maiden and NewDISS Team 2000). The recommendations from the NewDISS Strategy Team are quoted below from the Executive Summary of the report by Maiden et al (2000):

- Support a spectrum of heterogeneous technological approaches to NewDISS. This includes concentrating on integrating suitable existing data service capabilities, while also identifying and providing a means for delivering capabilities that do not yet exist.
- Clearly define the components of NewDISS, and ensure suitable management of the interfaces between them. This includes the definition of a set of "core" standards and practices, along with the means for selecting and maintaining them.
- Employ a NewDISS infrastructure that includes active liaison with service providers both within NASA and within the private sector for procurement of common operations activities.
- Employ competition and peer review in the process used for choosing NewDISS components.
- Empower science investigators with an appropriate degree of responsibility and authority for NewDISS data system development, processing, archiving and distribution.

- Use lessons learned from the current, experimental ESE federation as a step towards the NewDISS, and proceed with the Federation Experiment with this evolution in mind.
- Charter, without delay, a transition team with the objective of developing a transition plan, based on the findings and recommendations of this document that would lead to the initiation of a NewDISS starting in 2001.

Addressing these recommendations, NASA initiated a formulation study called Strategic Evolution of Earth Science Enterprise (ESE) Data Systems (SEEDS) during 2002–2003. NASA's GSFC conducted this study with significant involvement by the scientific user community. The focus of this study was on how a system of highly distributed providers of data and services could be put in place with community-based processes and be managed by NASA. The areas considered in this study were: levels of service and costs, near-term mission standards, standards and interfaces processes, data life cycle and long-term archive, reference architectures and software reuse, technology infusion, and metrics planning and reporting. As a result of the recommendations from this study, NASA established a set of four Earth Science Data System Working Groups (ESDSWG): Standards Processes, Reuse, Technology Infusion, and Metrics Planning and Reporting. The EDSWG continues to meet in groups and in plenary to promote integration of standards and capabilities into NASA's Earth Science Systems.

NASA views its data systems in terms of "Core" and "Community" capabilities. The core capabilities provide the basic infrastructure for robust and reliable data capture, processing, archiving, and distributing a set of data products to a large and diverse user community. Examples of core capabilities are: 1. The Earth Observing Data and Information System (EOSDIS); 2. The Precipitation Processing System (Stocker 2003); 3. Ocean Data Processing System (Feldman 2007); and 4. The CloudSat Data Processing Center (NASA and CSU 2007). EOSDIS is a multimission data system that manages data from all of the EOS missions and most of the heritage (pre-EOS) missions. The Precipitation Processing System is recently evolving as a measurement-based system from the Tropical Rainfall Mapping Mission Science Data and Information System (TSDIS) and is planned to support data management for the Global Precipitation Mission (GPM). The Ocean Data Processing System, managed by the Ocean Biology Processing Group at NASA GSFC, is a measurement-based system that spans several missions ranging from Nimbus-7 to EOS. The CloudSat Data Processing Center is a system specific to CloudSat, one of the missions in the Earth System Science Pathfinder (ESSP) program. The latter three examples are "loosely coupled" with EOSDIS, in that they exchange data with the EOSDIS Data Centers and are consistent with EOSDIS in the use of data format standards.

In contrast to the core capabilities, community capabilities provide specialized and innovative services to data users and/or research products offering new scientific insight. Such systems are generally supported by NASA through peer-reviewed competition. Examples of community capabilities are projects under the Research, Education and Applications Solutions Network (REASoN), Advancing

Collaborative Connections for Earth System Science (ACCESS), and Making Earth Science Data Records for Use in Research Environments (MEaSUREs) Programs.

Both core and community capabilities are required for NASA to meet its overall mission objectives. The focus of the ESDSWGs is on community capabilities. While the membership on the four working groups is open to all, the primary participation is by members of the REASoN, ACCESS, and MEaSUREs projects. The working groups are a mechanism through which the community provides inputs for NASA to help with decisions relating to Earth science data systems. There is significant commonality in membership between the ESDSWG and ESIP Federation, thus bringing into the NASA Earth science data systems a broad community perspective.

5.4 Use of Standards and Interfaces in EOSDIS

The development and use of standards within the EOSDIS architecture has been one of the real success stories of the ESDIS Project. Standards play a critical role in how EOSDIS will serve to meet future needs. By adopting standards, we hope to foster inter-organizational data discovery and manipulation. To be useful and effective, standards must always be reviewed and modified. The ESDIS Project has always made a resource commitment to maintain and develop standards. The ESDIS Project has also opened the doors to the greater community by providing mechanisms to discuss and integrate standards into the EOSDIS. Early adoption of community standards by EOSDIS has proved to be a cost benefit to the ESDIS Project by allowing easier integration of new missions into the baseline, by reducing the complexity of the system, by reusing existing software and processes, and by enabling easier cross-training across EOSDIS.

EOSDIS has several ongoing standards activities. These include:

- Direct standards such as data format and metadata standards
- Standard usage of terms and documentation
- Standardized processes and metrics

5.4.1 Data Formats

EOSDIS has fostered the development of several standards used within science data processing systems. Historically, the format of data products was picked by the principal investigators of each individual science instrument based on convenience and cost benefit to the processing teams. In order to facilitate the ability for diverse communities to use data in interdisciplinary studies, early in the development of EOSDIS the ESDIS Project conducted a collaborative study with the EOSDIS Data Centers of the then available standard formats for adoption in EOSDIS. None of these formats met all of the requirements. However, the Hierarchical Data Format (HDF), developed by the National Center for Supercomputing Applications (NCSA)

at University of Illinois, satisfied most of the requirements. Therefore, the ESDIS Project selected this to be the data format (actually, a data formatting system with associated software tools) to be used for archiving and distributing data products for EOS instruments. The ESDIS Project has been supporting the maintenance and evolution of this formatting system, first at the NCSA and later at the HDF Group (THG), a non-profit corporation. The EOSDIS Data Centers also maintain heritage data in other (native) formats, and provide format translations to users as needed.

The HDF is a multi-object file format that facilitates transfer and manipulation of scientific data across multiple systems. It supports a variety of data types. The HDF library provides a number of interfaces for storing and retrieving these data types in compressed or uncompressed formats. HDF files are self-describing and permit users to understand the file structures from information stored in the file itself. However, the traditional HDF file structure does not include geolocation information. Since it is critical for Earth observation data to be geolocated, the ESDIS Project developed the HDF-EOS format that included additional conventions and data types for HDF files. The three geospatial data types supported by HDF-EOS are Point, Grid, and Swath. Using the HDF format, the ESDIS Project took an additional step to identify three ways of looking at EOS instrument data products: point products, grid products and swath products. The standard HDF tools can also read HDF-EOS files. However, the HDF-EOS library provides software for easier access to geolocation data, time data, and product metadata than the standard HDF library.

A key feature of these three product types in HDF-EOS is the identification of core metadata values that must accompany all products for inclusion into EOSDIS. Each of the EOS science data teams is required to submit data in the HDF-EOS format, with waivers provided only where justifiable. The Project provides many avenues of assistance to facilitate the acceptance of this standard by users including user guides, specialized software libraries, forums, websites, and a yearly HDF Workshop held in areas across the US. Because the data format is widely published, the community is able to propose and develop tools to read and manipulate EOSDIS data. Two types of the most popular tools are subsetters and reprojection tools. More details on HDF and HDF-EOS can be found on the HDF web site (HDF 2008).

5.4.2 Metadata Standards

EOSDIS has a strong commitment to metadata standards. Twenty years ago, the concept of deriving metadata from the actual data was considered burdensome to the science data producer community. Despite this initial resistance, the ESDIS Project created the EOSDIS Core data model, which describes a standard set of metadata that are required for each data collection and products within the collection. Standard metadata required from the data providers include such basic information as product name, type, collection information, time of acquisition, and geographic coverage. The core data model was developed while the US Federal Geographic

Data Committee (FGDC) was developing the metadata content standard to be followed by US agencies. The extensions of the FGDC standard for remote sensing metadata have been influenced by the EOSDIS Core data model (FGDC 2002).

This basic requirement has served to enable the development of a rich set of user interfaces and data discovery tools. All EOSDIS metadata are accessible through the EOSDIS Clearing House (ECHO) interface. ECHO has public application programming interfaces that allow access to the metadata, which are published in XML format. Use of the XML format is another standard adopted that allows for easy access by all types of World Wide Web interfaces.

EOSDIS Data Centers also use the Open Geospatial Consortium (OGC) web services. OGC standards have a particular affinity to geolocated data and are beneficial to users of many data products offered by EOSDIS. EOSDIS is starting to implement two particular web services: the Web Mapping Service (WMS) and the Web Coverage Service (WCS). As more data at the EOSDIS Data Centers are made available in WMS/WCS, users will be able to layer many types of NASA data on geospatial information systems. Use of Google KML files to layer EOS data on Google Earth is another standard that EOSDIS is adopting.

5.4.3 Terms and Documentation Standards

No discussion of standards within EOSDIS would be complete without a discussion of the usage of standardized terms and documentation. For example, the term "granule" was adopted early by the ESDIS Project to mean the smallest instance of a data product tracked in the database for searching, ordering, and/or access. A granule can be one or more files. The concept of the granule is now universally understood within NASA Earth science communities.

The term "browse" is another standard fostered in EOSDIS. "Browse" data is now commonly understood to be small thumbnail images of the actual data. Access to browse data enables users to examine the dataset for desired features prior to the potential time-consuming step of downloading large datasets.

The ESDIS Project also focused on providing standard approaches to documentation. Every data collection in the system includes the Directory Interchange Format (DIF) registration, which enables search from the NASA Global Change Master Directory (GCMD). Collections also include a standard guide document to the data set.

5.4.4 Process Standards

The ESDIS Project has developed several standardized processes to facilitate the configuration control and the management of the EOSDIS. Standardized processes are uniformly applied across EOSDIS elements. Data processing teams at the SIPSs and EOSDIS Data Centers participate in preparing and reviewing interface control documents and other related documentation. The ESDIS Project established the

management tools and processes early in the ESDIS Project lifecycle to apply a routine approach to reviewing and changing documentation associated with EOSDIS. All project plans, requirements, and interface control documents are accessible on the ESDIS Project web pages.

Capturing system performance metrics is another example of a uniform process applied across the elements of the EOSDIS. Metrics such as product distribution, archive size, and data center web activity are defined and reviewed at the ESDIS Project level, and each data center provides a standard set of measurement inputs to the Project. Common metrics are then available not only to ESDIS management, but also to the metrics providers. Better project management is enabled by allowing the data centers access to their detailed metrics, at the same time allowing the ESDIS Project to have a system view across all of EOSDIS.

5.4.5 Standards to be Developed

While the EOSDIS has made great progress toward the introduction and common usage of standards, areas for improvement exist. We would like to see the development of "provenance" standards to provide the ESDIS Project more complete information concerning the source and make-up of datasets. Provenance standards include the identification of information needed for the long-term archive of datasets and associated material (e.g., documentation). The need for provenance standards is critical to establish both the heritage and quality assessment of the data.

Another area where standards still need attention is in the selection of dataset map projections. Despite efforts at coordination in the early years of the EOS mission design, each science instrument team was allowed by the EOS Program to determine the best projection and scale to be used for its data. Consequently, many differing projections are used for EOSDIS data. This makes it difficult for users to integrate and inter-use data from multiple instruments or disciplines.

5.5 Evolution of EOSDIS Elements Study

In late 2004, NASA Headquarters management initiated a review of the EOSDIS. NASA prepared a charter for the "Evolution of EOSDIS Elements (EEE) Study" (Cleave 2004) with the goal to "assess, by considering the future objectives, the current state of EOSDIS in order to identify the components that can/must evolve, those components that need to be replaced because of the rapid evolution of information technologies, and those components that require a phase-out strategy because they are no longer needed." The charter advanced objectives for the study as:

- Increase end-to-end data system efficiency and operability
- Increase data usability by the science research, application, and modeling communities

- Provide services and tools needed to enable ready use of NASA's Earth science data in the next-decadal models, research results, and decision support system benchmarking
- Improve support for end users

The EEE charter established two teams to accomplish these goals: a Study Team and a Technical Team. The Study Team received direction to provide recommendations consistent with the goal and objectives stated above and was charged with looking at the existing EOSDIS to determine the strategic evolution of its functions and elements in the broader context of the processes, goals, and objectives of the NASA Earth science strategy and plans for the next decade's data systems and architectures. The Study Team consisted of nationally recognized technical experts in Earth system science, applications, and information technology. The Technical Team, led by the ESDIS Project Manager, was made up of representatives from the ESDIS Project, DAACs, and SIPSs, and selected consultants invited to provide independent perspectives on aspects of data system development from their experiences. The Study Team, along with the Technical Team, prepared a vision for the Evolution of EOSDIS Elements, (EEE Study Team 2005) projecting the system capabilities to the year 2015. The Vision emphasized the need to ensure safe stewardship of the Earth science data while maintaining technological currency to further enable scientific research based on EOSDIS data holdings. This Vision provided the guiding principles under which the Technical Team conducted its analytical work. The goals expressed in the Vision and the tenets derived from them by the Technical Team are shown in Table 5.2. These goals and tenets were used in tracking the progress of evolution towards the Vision.

The Technical Team performed a detailed analysis of the EOSDIS components and elements and developed an approach and implementation plan that would begin to fulfill the objectives set forth in the Vision. The Technical Team sought inputs from the lead individuals of each of the current system elements and encouraged their contribution of ideas and concepts for improving EOSDIS consistent with the Vision. Selected consultants were invited to provide independent perspectives on aspects of data system development from their experiences.

The Technical Team analyzed the suggestions for: adherence to the Vision and Study Team guidance; the investment costs, sustaining costs, and lifecycle costs; identification of the potential risks; implementation feasibility; timeframes and phasing opportunities; and for the effect on the user community. With this analysis, the element inputs were structured into the following set of alternative approaches:

- DAAC-focused – all DAACs develop their own archive management systems to reduce dependence on the core systems,
- SIPSs-focused – the SIPSs take on the archive, distribution and customer interface responsibilities in place of the DAACs, and
- Core System-focused – implement a re-architected ECS at all four DAAC sites where it was deployed at that time.

Table 5.2 EOSDIS evolution vision – tenets and goals

Vision tenet	Vision 2015 goals
Archive management	• NASA will ensure safe stewardship of the data through its lifetime. • The EOS archive holdings are regularly peer reviewed for scientific merit.
EOS data interoperability	• Multiple data and metadata streams can be seamlessly combined. • Research and value added communities use EOS data interoperably with other relevant data and systems. • Processing and data are mobile.
Future data access and processing	• Data access latency is no longer an impediment. • Physical location of data storage is irrelevant. • Finding data is based on common search engines. • Services invoked by machine-machine interfaces. • Custom processing provides only the data needed, the way needed. • Open interfaces and best practice standard protocols universally employed.
Data pedigree	• Mechanisms to collect and preserve the pedigree of derived data products are readily available.
Cost control	• Data systems evolve into components that allow a fine-grained control over cost drivers.
User community support	• Expert knowledge is readily accessible to enable researchers to understand and use the data. • Community feedback directly to those responsible for a given system element.
IT currency	• Access to all EOS data through services at least as rich as any contemporary science information system.

From these three alternatives, a hybrid approach was defined, selecting the best aspects of each alternative that could be feasibly developed in concert. This fourth alternative, the Hybrid Approach, was advanced as the "best value" for cost containment, risk management, and fulfillment of Vision goals. NASA Headquarters approved this Hybrid Approach and directed the Technical Team to plan for its implementation.

5.6 Implementation of the Evolution Plan

The Hybrid Approach for implementation involved activities in five major areas of EOSDIS. These activities were carried out in parallel during the years 2006–2008. The first activity was re-architecting of ECS, consisting of simplifying the software and hardware architectures to reduce maintenance costs while improving service. This simplified ECS was deployed at three of the four data centers (Langley ASDC, LP DAAC, and NSIDC). The second was the addition of the archiving and distribution functions for MODIS Level 1 and atmospheric data products

to the MODIS Adaptive Processing System (MODAPS) SIPS, using on-line disks for the archive and reducing the size of the archive by processing Level 1 products on demand when deemed advantageous. The third was the deployment of the Simple Scalable Script-based Science processor Archive (S4PA) system to replace the current ECS system at the GES DISC. Along with this development, all the data at the GES DISC were made accessible on line. The fourth was the development of the Archive Next Generation (ANGe) at the ASDC at LaRC. This replaced the Langley TRMM Information System (LaTIS) which had been used for processing, archiving, and distributing CERES data from TRMM, Terra, and Aqua missions. Also, the processing system for the Terra Mission Multi-Angle Imaging Spectrometer (MISR) instrument was migrated to a Linux cluster. The fifth activity was the completion and deployment of the EOS Clearing House (ECHO) as a robust operational middleware system.

Each of these activities will be discussed briefly in the following subsections.

5.6.1 ECS Re-architecting

In 2005, the EOSDIS Core System (ECS) was deployed at four EOSDIS Data Centers that perform ingest, processing, archive and distribution of EOS data. The system architecture consisted of two loosely coupled systems: the original Science Data Processing System (SDPS) which provided ingest, archive, processing, and distribution functions utilizing a large scale, multi-petabyte tape archive, and the newer Data Pool which provided data access and distribution via a large, shared disk store. The design of the original SDPS (circa 1995) was oriented towards the complexities of managing a tape archive, limited computing resources (CPU, memory, small direct-attached caches), and a sophisticated, type-extensible data model. The system was complex and large (over 1 million lines of custom C++ code plus scripts and database-stored procedures). The SDPS hardware architecture was based on enterprise-class SGI and Sun UNIX servers with direct attached storage and host-centric file systems. This hardware suite became expensive to maintain, and required custom code to be supported on two different operating systems (IRIX and Solaris). The Data Pool design leveraged new technology and experience with actual EOSDIS operations, utilizing a simplified data model, Order Management services based on the data being available online, and a hardware architecture built around a Storage Area Network (SAN).

The Evolution approach for ECS has many new features affecting the software, hardware, and maintenance processes.

Data ingest and distribution functions were re-implemented as Data Pool services. This enabled retirement of the legacy Storage Management and Data Distribution subsystems. User search and order functions were allocated to the evolving ECHO infrastructure, enabling retirement of SDPS user support tools and gateways. The complex Science Data Server database and custom software were replaced with a greatly streamlined Archive Inventory Management database. All metadata are now stored in XML. The result of these changes was a significant

reduction in the custom code to be maintained, from 1.2 million source lines of code (SLOC) to approximately 400 K SLOC.

At each data center, all custom code applications and commercial-off-the-shelf software are now running on new hardware as a single blade cluster. All storage is now provided by the SAN, eliminating most of the network data transfers between hardware platforms. The Data Pool SAN capacity was increased to accommodate ingest buffers, and increased storage and distribution capacity. Databases were consolidated onto a single Linux-based database server at each data center.

The re-architected ECS introduced many changes to its suite of commercial-off-the-shelf software. The Hierarchical Storage Manager product was replaced with a newer product that is supported on Linux, runs on commodity hardware platforms, and reduces the archive administration workload on operators. This also enabled a significant reduction in licensing costs, and set the stage for the eventual migration to a totally on-line (disk-based) archive. Operating system maintenance has been simplified by reducing to one (i.e., Linux).

Experience in operating the ECS for 10 years identified many opportunities for improving operations efficiency through automation of operator-intensive tasks. This included automation of recovery from failure conditions; improved operator interfaces to simplify operations; and automation of human-intensive data management tasks.

The above evolution was implemented in a phased approach to minimize impact to operations. The new hardware architecture was deployed to the data centers while operations continued on the legacy configuration. All custom code was ported to run on the Linux operating system. The initial software release (which transitioned data ingest and distribution functions to Data Pool services) was implemented so that it could run on both the new Linux-based, commodity hardware architecture, and the legacy hardware. After the new software architecture was stable on the new hardware architecture, the legacy hardware was removed. The development and test facilities were transitioned first (in 2006) and the EOSDIS Data Centers migrated to the blade/SAN architecture in 2007. The second major software release, which enabled retirement of the legacy Science Data Server and transitioned all user search and order functionality to ECHO, occurred in mid-2008.

The goal of reducing recurring costs was achieved. With a reduction of 33% of the annual ECS operations and system maintenance costs, the investment in new hardware and software was repaid in the first year of the Evolution activity.

5.6.2 LAADS/MODAPS

The MODIS Adaptive Processing System produces the base level (i.e., Level 1), land, and atmosphere data products from the MODIS science instrument data. Before the Evolution activity, MODAPS served as an EOSDIS SIPS (See Sect. 5.2.3) and the GES DISC processed Level 1 MODIS products, and archived and distributed the Level 1 and atmospheric products. Land products were archived

and distributed by LP DAAC and the Snow and Ice products by NSIDC. EOSDIS Evolution resulted in migration of the archiving and distribution functions for MODIS Level 1 and atmospheric data from the GES DISC to MODAPS. By incorporating the MODIS data archiving and distribution functions into MODAPS the EOSDIS gained much efficiency. The part of MODAPS that performs these functions is referred to as the Level 1 and Atmospheric data Archiving and Distribution System (LAADS). All the other functions for MODIS data have remained at the same EOSDIS elements as indicated above.

The archive at MODAPS was planned as a fully disk-based archive. The Level 1 products that contributed a high production volume (54% of daily production) would be held on line for a short period for users to download and would be produced on demand after that initial period. Benefits of this transition included:

- Reduction in archive growth through on-demand processing
- Faster access to products, reduced reprocessing time from all on-line storage
- Reduced costs due to use of commodity disks and simplification of operations
- Closer involvement and control by the science community, greater responsiveness to scientific needs, products, tools, and processing

Details about the implementation of LAADS can be found in Masuoka et al. (2007). A few key points are mentioned below.

The foundation for LAADS is the MODAPS processing software. This software is now augmented with features that support product search, ordering, and distribution. The production systems send the generated data product files to LAADS using scripts for ingestion. As a part of the ingest process, the LAADS archive database is updated. Users can search for data based on spatial and temporal criteria as well as quality thresholds. Inexpensive high-capacity disk drives have enabled MODAPS to significantly increase the processing rates and to improve distribution of Level 0 data as well as higher-level data products to the user community by storing all except Level 1 products on line in the disk archive. Large products, produced early in the MODIS processing chain, such as the Level 1A (unpacked instrument counts) and Level 1B (calibrated radiances) are produced on demand. Such products may be regarded as "virtual products".

MODAPS also offers options including reprojection, masking, subsetting, and reformatting for transforming both archived and virtual products into forms better suited for an individual user's needs. When a user requests an on-demand or custom product, jobs are launched on a cluster of compute servers assigned to LAADS and the results placed in an on-line directory for retrieval by the requestor. The products generated on demand are held in the on-line archive for a period of a few days to enable the requesting user (and any other interested user) to download the files. Also, with even further reductions in on-line storage costs, it is becoming feasible to hold larger percentages of the Level 1 products on line thus reducing the need for on-demand production.

The products in the on-line archive can be obtained by users through a variety of means. Interactive search and order can be performed through the LAADS web

server. A file transfer protocol (FTP) server is available for scriptable access to the entire archive. Some servers support machine-to-machine access protocols. Also, users can perform cross-instrument and cross-data center searches and order products through ECHO and the Warehouse Inventory Search Tool (WIST). Of these access methods, the most popular means of obtaining data products is via FTP.

The MODAPS evolution and the implementation and transition to operations of LAADS from GES DISC occurred as a gradual process during February 2006 and January 2007 to ensure no disruption to the users. Before the transition, the data were archived in robotic tape silos. The average monthly distribution during the 6 months prior to transition was 3 million files and 30 terabytes per month. Soon after the transition, with all data on line, these numbers increased to 7 million files and 48 terabytes, respectively, per month. In late 2008, these numbers were at over 150 million files and 80 terabytes per month.

5.6.3 GES DISC

From its inception as one of the original EOSDIS DAACs the Goddard Earth Sciences (GES) Data Information Services Center (DISC), developed and managed two data management systems concurrently. The first was the Version 0 system for TRMM and pre-EOS heritage data. The second system was the EOSDIS Core System used for NASA's Terra, Aqua, and Aura missions. The evolution plan at the GES DISC was to reduce operations to a single data management system.

The GES DISC evolution was based on an in-house developed software system and adoption of on-line storage using commodity based hardware. The data center staff initially developed the Simple, Scalable, Script-based Science processor Archive (S4PA) system to replace the V0 system. The S4PA is a simplified data archive architecture where data resides on commodity disks. Its modular design permits its reuse with replacement of application-specific components. The transitions from ECS to S4PA were handled incrementally to ensure no impact to ongoing operations. Given the reduction in volumes due to transfer of MODIS data archiving to LAADS/MODAPS as indicated above, the ECS robotic silos were phased out. This represents a cost savings to the GES DISC and to the EOSDIS Core System as well.

The benefits of these changes included:

- Reduced operations costs due to consolidation of multiple systems into one software system
- Increased data center automation due to a single management system with simpler operational scenarios
- Reduced sustaining engineering costs due to use of simpler, scalable software and reduction in dependency on high maintenance commercial-off-the-shelf products
- Improved data access due to increased on-line storage and commodity disks/platforms
- Risk mitigation for the LAADS/MODAPS transition effort

The size of the GES DISC archive was significantly reduced with the transition of MODIS data management responsibility to MODAPS. The remaining science data at the GES DISC was archived on line, eliminating the need for tape silo storage. By the end of 2007, the GES DISC completed migration of data to S4PA.

With the data stored on line, users have greater flexibility for access to data. Users can navigate to the data of interest through the hierarchical structure of S4PA or write scripts to acquire bulk data. GES DISC also offers services such as OPeNDAP (Cornillon et al. 2003), OGC Web Map Service and Web Coverage Service, and on-line analysis capabilities using Giovanni, which is a Web-based application developed by the GES DISC that provides capabilities to visualize, analyze, and access Earth science remote sensing data without having to download the data (Acker and Leptoukh 2007). Users can search for data by navigating through the hierarchical structure as indicated above, a web based hierarchical navigation tool or a free-text (Google-like) tool called Mirador. The GES DISC also supports cross-data center searches through the WIST client using the ECHO middleware. Since the data are archived on disks, GES DISC can tailor its services to particular missions or measurements and provide discipline specific services. More details about the evolution of GES DISC can be found in a paper (Kempler et al. 2009).

5.6.4 Langley ASDC

The Langley Atmospheric Science Data Center, an original EOSDIS DAAC, employed two data management systems over time to meet its needs for processing, archiving, and distribution functions. The first, called the Langley Tropical Rainfall Measuring Mission (TRMM) Information System (LaTIS), was used for processing, archiving, and distributing the Clouds and the Earth's Radiant Energy System (CERES) instrument data products and to archive and distribute all the pre-EOS data products held at this data center. The second system, the ECS, was used for processing, archiving, and distributing the data products from the Multi-angle Imaging Spectro-Radiometer (MISR) instrument on the EOS Terra spacecraft, and for archiving and distributing data products from several other EOS instruments.

The ASDC's evolution strategy consisted of replacing LaTIS with a modern, advanced, scalable system developed in-house called Archive Next Generation (ANGe) (Ferebee et al. 2007). ANGe is designed to increase automation and reduce manual operations in the archiving and distribution functions, and be expandable for additional science data. ASDC also upgraded their ECS implementation with the re-architected version of ECS (described in Sect. 5.6.1 above). The benefits of these changes included:

- Reduction in sustaining engineering costs due to reduction in dependency on high maintenance commercial-off-the-shelf products
- Increased system automation
- Improved data access due to planned use of increased on-line storage and commodity disks/platforms

ANGe began to successfully archive and distribute CERES and Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) datasets in 2008. In addition to ANGe development at the ASDC, the software for processing MISR data has been migrated from the SGI systems to Linux clusters to increase efficiency and scalability and to reduce maintenance costs.

5.6.5 ECHO

The development of the EOS Clearing House (ECHO) was underway before the Evolution activity began, but it incorporated enhancements to meet Evolution objectives. ECHO is implemented as a series of releases adding or enhancing capabilities. The extension of ECHO capabilities is important for the EOSDIS Evolution to meet the objectives of data interoperability, data access, and preserving the pedigree of derived data products. Since the EOSDIS Evolution activity began, ECHO has added capabilities, including Collection and Granule Browse data insert, update, and delete; Enhancements to Access Control Lists (to support multiple collections per provider); Mechanism for Clients to perform Spatial Query based on Latitude/Longitude; Line Item Order Status; Reorganization of Web Services Applications Program Interface (API) to improve usability; Framework for error handling; and changes to improve maintainability and performance. Additional functionality under development includes support for asynchronous queries by ECHO Clients, support for Product Specific Attributes, and support for 2-dimensional coordinate-based search (e.g., path/row).

Future versions will include a new Web Service Order interface; new capabilities in the areas of metrics reporting, event notification, and data partner data reconciliation; improved performance from metadata transmission to ECHO ingest; and improvements to better ensure data integrity.

5.7 Progress Towards Vision 2015

The Evolution planning teams characterized the Vision for 2015 in seven tenets, each representing a set of objectives guiding evolution success. These tenet goals, presented in Table 5.2, provide a mechanism for gauging the results to date. After nearly three years, the implementation activity discussed in Sect. 5.6 shows significant progress in meeting the goals of each Vision tenet.

The evolution of EOSDIS resulted in progress towards the Vision for 2015 by implementing changes that maximize science value and achieve cost savings. At the current stage in its evolution, EOSDIS makes data access easier and data products more quickly available to the science community by increasing the amount of data available on line. EOSDIS data has also become more closely integrated with the science community, especially with MODIS data for the atmospheres community. Substantial cost savings have been achieved by replacing operations with automation, seeking less costly sustaining engineering approaches, and taking advantage of current information technology advances in hardware and automation.

Progress towards each Vision tenet is discussed below.

Archive Management has been strengthened through upgrades to the hardware and software at each site. More data is available on line. The LAADS/MODAPS data center and the GES DISC have evolved away from tape archive based systems and the re-architected ECS at other data centers has increased the use of data pools. EOSDIS data centers have tailored their processing and archiving software and systems to be more efficient. The data centers now have the ability to review the archive collection. This ability to better manage the archive supports the goal of long-term data stewardship.

EOS Data Interoperability is enhanced by making more data available on line, thus decreasing the access time to the science data and products.

The ease of access to EOSDIS data is evidenced through a dramatic increase in data distribution to end users. EOSDIS has experienced increases in product distribution of approximately 50% in both FY2007 and FY2008. In FY2006, the number of products distributed increased by less than 20% over the previous year. While some of this increase may be the result of a general worldwide growing interest in Earth science-related problems (e.g., climate change), the magnitude of the increase suggests that data are becoming easier to access.

Other aspects of interoperability are needed to achieve the 2015 Vision. EOSDIS can now focus on defining ways for combining multiple data and metadata streams seamlessly, and can address data interoperability with other relevant data systems. While on-line availability facilitates making processing and data mobile, it takes more effort and coordination with the science community to achieve this fully.

Future Data Access and Processing objectives are being met by archiving data on line and processing on demand, which support provision of services that customize data access in the amounts and the form needed by science users. Because of the vast increase in processor speed, EOSDIS is now able to process on demand. The ability to process on demand also reduces the size of the archive to be maintained. For example, EOSDIS ingests a large amount of data from the MODIS instrument alone, which progresses through multiple levels of processing to become useful products. At one time storing the initial processed (Level 1) data required a large archive for data that were not highly sought after nor uniformly requested from across the entire global coverage area. By not archiving these intermediate data products but ensuring availability by reprocessing lower level products as needed, EOSDIS saves storage space at the modest cost of some reprocessing. Also, the design of the evolved EOSDIS permits more agile decisions on processing versus storage based on changes in hardware technology and the resulting reductions in cost.

The ECHO middleware provides a robust and common means to access EOS data. Beginning in June 2008 EOSDIS Data Centers began transitioning from the EOSDIS legacy user interface system (EDG) to the EOSDIS Warehouse Inventory Search Tool (WIST) system. From a general user perspective, the access to data depends little on where it is physically located, or even the means to prepare it for delivery, as long as the data are made available in a reasonable amount of time.

The ability to track the *Data Pedigree* improves with the focus on metadata and the success of the evolving EOSDIS Core data model. More attention is needed

for preserving and ensuring access to the various versions of the software used to generate the data products.

Cost Control was improved with a focus on identifying and evolving the components that were cost drivers. All data center sites began a process to transition from expensive workstations to commodity hardware and the replacement of expensive commercial-off-the-shelf tools with less expensive tools while retaining essential functionality. This included new maintenance strategies (e.g., purchase of less costly computing platforms with a more frequent refresh cycle rather than paying high maintenance costs); and increased automation leading to operational cost savings. Targeting specific data centers for initial improvements produced earlier and larger cost savings. The other data centers followed with self-directed upgrades to equipment replacement, software upgrades, and archive holdings' cleanup in parallel with the formal evolution process.

User Community Support improved by moving control of the data and supporting services closer to the users and science teams. A specific example from the EOSDIS Evolution is the support to the atmospheres community by combining the MODIS archive and distribution with the data processing function. This closer tie between the user community and data providers enables EOSDIS to be responsive to science requirements, and influences the product definition, tool development, and processing needs to the benefit of the users. The EOSDIS Data Centers have increased the number of on-line tools and services, such as visualization and sub-setting. All user communities, including the general public, are served by improved interfaces, the upgraded catalog and inventory tools, and the easier access to data on line.

Information Technology (IT) Currency is being realized in the upgrades and simplifications provided by the re-architected ECS and the data center upgrades, improving the flexibility to meet expectations of a more sophisticated user community. The entire EOSDIS Evolution of Elements activity is an example of the NASA commitment for continuous technology assessment and infusion.

5.8 Summary

In this chapter, we have provided a discussion of the evolution of EOSDIS. As NASA's major data system capability for managing Earth science data, EOSDIS has been evolving since its conception in the early 1990s. Many changes have occurred along the way. Starting with a centralized design involving two data centers, it was changed to have a more geographically distributed set of eight Distributed Active Archive Centers (DAACs), where each was focused on a specific set of Earth science disciplines. The design of the system where all the EOS standard products were to be generated at the data centers using the EOSDIS Core System and instrument team provided software was replaced by an implementation using Science Investigator-led Processing Systems (SIPSs) to generate the data products. Currently there are twelve EOSDIS Data Centers and fourteen SIPSs in EOSDIS. The standards, discussed in Sect. 5.4 of this chapter, have played an important role in the successful operation of these elements as well as in interactions with the user community.

The latest focused effort for evolution of EOSDIS started as a formal study initiated by NASA in 2004 with the goals of increasing efficiency, operability, data usability, services and tools availability, and improved support for end users. The Study Team and the Technical Team working on this evolution study arrived at a Vision for 2015, with the Technical Team following up with an initial implementation plan. This implementation plan was established by late 2005, and the implementation was carried out during 2006 through 2008. The major elements involved in this implementation were the EOSDIS Core System, MODAPS, the GES DISC, the Langley ASDC, and ECHO. The simplified, re-architected ECS is now operating at three EOSDIS Data Centers – Langley ASDC, LP DAAC, and NSIDC. (The remaining Data Centers do not depend on ECS for their operations). As of this writing, most of the implementation has been completed, and the systems are in operation. Significant progress has been made towards the Vision for 2015 goals.

Despite all of this progress, much work remains to ensure that EOSDIS remains a vibrant tool to serve the user community as technology changes over the next several years. The rapid changes in information technology will provide both challenges and opportunities. The challenges will be due to increased expectations on the part of users concerning innovative uses of data and information to derive knowledge. The changes in technology will provide opportunities for EOSDIS and its elements to improve their capabilities to serve the community in innovative ways. A strategy for active infusion of technology is essential for the continued success of EOSDIS.

Acknowledgments The development and evolution of EOSDIS since 1990 has been the result of the efforts of scores of people within and outside NASA. It is not practical to name all those that have been involved. The most recent, formal study was conducted as a collaborative effort between the EOSDIS Elements Evolution Study Team and the Technical Team, further supported by technical consultants. The Study Team members were M. Pniel (lead), W. Brooks, P. Cornillon, S. Denning, J. Frew, W. Green, and B. Minster. M. Maiden and M. Esfandiari were ex officio members. The Technical team members were: M. Esfandiari (lead), J. Behnke, C. Bock, M. Ferebee, K. Fontaine, M. Goodman, P. Liggett, D. Lowe, K. McDonald, E. Masuoka, D. Marinelli, K. Moe, R. Pfister, H. Ramapriyan, S. Reber, C. Schroeder, E. Sofinowski, S. Turner, and B. Vollmer. The technical consultants were G. Feldman, C. Lynnes, E. Stocker, and V. Zlotnicki. The implementation was carried out by the ESDIS Project and each of the EOSDIS elements mentioned in Sect. 5.6.

Acronyms

ACCESS	Advancing Collaborative Connections for Earth System Science
ACRIM	Active Cavity Radiometer Irradiance Monitor
ACRIMSAT	Active Cavity Radiometer Irradiance Monitor Satellite
AIRS	Atmospheric Infrared Sounder
AMSR-E	Advanced Microwave Scanning Radiometer - EOS
AMSU	Advanced Microwave Sounding Unit
ANGe	Archive Next Generation
API	Application Programming Interface
ASDC	Atmospheric Sciences Data Center
ASF	Alaska Satellite Facility
ASTER	Advanced Spaceborne Thermal Emission and Reflection

CALIOP	Cloud-Aerosl Lidar with Orthogonal Polarization
CALIPSO	Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation
CDDIS	Crustal Dynamics Data and Information System
CERES	Clouds and the Earth's Radiant Energy System
CPR	Cloud Profiling Radar
DAAC	Distributed Active Archive Center
DB	Direct Broadcast
DIF	Directory Interchange Format
DISC	Data and Information Services Center
DORIS	Doppler Orbitography and Radiopositioning Integrated By Satellite
ECHO	EOS ClearingHOuse
ECS	EOSDIS Core System
EDC	EROS Data Center
EDG	EOS Data Gateway
EDOS	EOS Data and Operations System
EEE	Evolution of EOSDIS Elements
EOC	EOS Operations Center
EOS	Earth Observing System
EOSDIS	Earth Observing System Data and Information System
EPGS	EOS Polar Ground Stations
EROS	Earth Resources Observation Systems
ESDIS	Earth Science Data and Information System
ESDSWG	Earth Science Data System Working Groups
ESE	Earth Science Enterprise
ESIPs	Earth Science Information Partners
ESSP	Earth System Science Pathfinder
FGDC	Federal Geographic Data Committee
FOS	Flight Operations Segment
FTP	File Transfer Protocol
GCMD	Global Change Master Directory
GES	GSFC Earth Sciences
GHRC	Global Hydrology Resource Center
GLAS	Geoscience Laser Altimeter System
GPM	Global Precipitation Mission
GSFC	Goddard Space Flight Center
HDF	Hierarchical Data Format
HIRDLS	High-Resolution Dynamics Limb Sounder
HSB	Humidity Sounder for Brazil
ICESat	Ice, Cloud and Land Elevation Satellite
IMS	Information Management System
IT	Information Technology
JMR	Jason Microwave Imager
JPL	Jet Propulsion Laboratory
LAADS	Level 1 and Atmospheric data Archiving and Distribution System

LaTIS	Langley TRMM Information System
LIS	Lightning Imaging Sensor
LP DAAC	Land Processes DAAC
MEaSUREs	Making Earth Science Data Records for Use in Research Environments
MISR	Multi-angle Imaging Spectrometer
MLS	Microwave Limb Sounder
MODAPS	MODIS Adaptive Processing System
MODIS	Moderate-Resolution Imaging Spectroradiometer
MOPITT	Measurements of Pollution in the Troposphere
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NCSA	National Center for Supercomputing Applications
NewDISS	new Data and Information Systems and Services
NISN	NASA Integrated Services Network
NOAA	National Oceanic and Atmospheric Administration
NRC	National Research Council
NSIDC	National Snow and Ice Data Center
OBPG	Ocean Biology Processing Group
OGC	Open GIS Consortium
OMI	Ozone Monitoring Instrument
ORNL	Oak Ridge National Laboratory
PB	Peta Byte
PI/TL	Principal Investigator/Team Leader
PO.DAAC	Physical Oceanography DAAC
QuickScat	Quick Scatterometer
REASoN	Research, Education and Applications Solutions Network
S4PA	Simple Scalable Script-Based Science Processor Archive
SAGE	Stratospheric Aerosol and Gas Experiment
SAN	Storage Area Network
SAR	Synthetic Aperture Radar
SDPS	Science Data Processing Segment
SeaWinds	Seawinds Scatterometer (For Flight On ADEOS II)
SEDAC	Socio-economic Data Applications Center
SEEDS	Strategic Evolution of Earth Science Enterprise (ESE) Data Systems
SIM	Spectral Irradiance Monitor
SIPSs	Science Investigator-led Processing Systems
SLOC	Source Lines of Code
SOLSTICE	Solar Stellar Irradiance Comparison Experiment
SORCE	Solar Radiation and Climate Experiment
TB	Terabyte
TDRS	Tracking and Data Relay Satellite
TES	Tropospheric Emission Spectrometer
THG	the HDF Group

TIM	Total Irradiance Monitor
TRMM	Tropical Rainfall Measuring Mission
TSDIS	Tropical Rainfall Mapping Mission Science Data and Information System
UARS	Upper Atmosphere Research Satellite
US	United States
USGS	US Geological Survey
UWG	Users Working Group
V0	Version 0
WCS	Web Coverage Service
WIST	Warehouse Inventory Search Tool
WMS	Web Mapping Service
WSC	White Sands Complex
WWW	World Wide Web
XPS	XUV Photometer System

References

Acker JG, Leptoukh G (2007) Online analysis enhances use of NASA earth science data. Eos, Trans., Am. Geophys. Union, 88 (2), 14, 17

Cleave ML (2004) Evolution of EOSDIS Elements Study Charter Amended. NASA http://eosdis-evolution.gsfc.nasa.gov/

Cornillon P, Gallagher J, Sgouros T (2003) OPeNDAP: Accessing data in a distributed, heterogeneous environment, Data Sci. J. [Online] 2 (0), pp. 164–174. http://www.jstage.jst.go.jp/article/dsj/2/0/2_164/_article

ECHO (2008) NASA EOS Clearing House. Online: http://www.echo.eos.nasa.gov/

EEE Study Team (2005) Evolution of EOSDIS Elements, Study Team Briefing to NASA. Online: http://eosdis-evolution.gsfc.nasa.gov/

EOS Project Science Office NASA GSFC (1990) EOS Reference Handbook

ESIPFED (2006) Federation of Earth Science Information Partners. Online: http://www.esipfed.org/

Feldman G (2007) Ocean Color Web. Online: http://oceancolor.gsfc.nasa.gov/, NASA (dynamically updated web page)

Ferebee MT, Cordner DE, Ritchey NA, Hunt LA, Piatko P, Haberer SJ, Wang FY (2007) Finding and accessing data at the NASA atmospheric science data center. IGARSS 2007, Barcelona

FGDC (2002) Content Standard for Digital Geospatial Metadata: Extensions for Remote Sensing Metadata. FGDC Document Number FGDC-STD-012-2002

HDF Group (2008) HDF-EOS Tools and Information Center. Online: http://hdfeos.net/index.php

Kempler SJ, Lynnes C, Vollmer B, Alcott G, Berrick S (2009) Evolution of Information Management at the GSFC Earth Sciences (GES) Data and Information Services Center (DISC): 2006–2007. IEEE TGARS, 21–28

Maiden ME, NewDISS Team (2000) NewDISS: A 6-to 10-year Approach to Data Systems and Services for NASA's Earth Science Enterprise, Draft, Version 1.0

Masuoka E, Wolfe R, Sinno S, Ye G, Teague M (2007) A Disk-Based System for Producing and Distributing Science Products from MODIS. IGARSS 2007, Boston, MA

NASA (2006) NASA Strategic Plan. NASA Headquarters, Washington, DC 20546, NP-2006-02-423-HQ

NASA, CSU (2007) CloudSat Data Processing Center. Online: http://www.cloudsat.cira.colostate.edu/

NASA GSFC, Science Advisory Panel for EOS Data and Information (1990) Panel Comments on EOSDIS (Phase B) Final Design Review, February 12–16, 1990

NASA GSFC (1989) UARS Ground Data Processing Description Document

National Research Council (1995) A Review of the US Global Change Research Program and NASA's Mission to Planet Earth/Earth Observing System. National Academy Press, Washington, DC

Ramapriyan HK (2003) NASA's Earth Science Data Systems – Past, Present and Future. IGARSS 2003, Toulouse, France

Ramapriyan HK, McConaughy GR (1991) Version 0 EOSDIS – An Overview. Technical Papers, ACSM-ASPRS Ann. Conv., 3, 352–362

Stocker EF (2003) A precipitation processing system for the Global Precipitation Measurement mission. Proc. Int. Geoscience and Remote Sensing Symposium. (IGARSS 2003), Toulouse, France, IEEE, 1704–1706

Chapter 6
SCOOP Data Management: A Standards-Based Distributed Information System for Coastal Data Management

Helen Conover, Marilyn Drewry, Sara Graves, Ken Keiser, Manil Maskey, Matt Smith, Philip Bogden, Luis Bermudez, and Joanne Bintz

6.1 Introduction

The Southeastern Universities Research Association (SURA) Coastal Ocean Observing and Prediction (SCOOP – http://scoop.sura.org) program is a SURA Coastal Research initiative that is deploying cutting edge IT to advance the science of environmental prediction and hazard planning for our nation's coasts. SCOOP is intended as a working prototype infrastructure to serve as a model for a distributed Integrated Ocean Observing System (IOOS) in the southeastern region (Bogden and Graves 2005). IOOS is a national initiative to create a new system for collecting and disseminating information about the oceans. The system will support a variety of practical applications, along with enabling research. A key partner in IOOS design and development, SURA is a consortium of over sixty universities across the US (http://www.sura.org). SURA's goal for SCOOP is to create a scalable, modular prediction system for storm surge and wind waves. Such networks will provide data in real-time and at high speed, for more reliable, accurate and timely information to help guide effective coastal stewardship, plan for extreme events, facilitate safe maritime operations, and support coastal security.

SCOOP is a distributed program, incorporating heterogeneous data, software, and hardware. This multi-institution collaboration is creating an open, integrated network of distributed sensors, data, and computer models to provide a broad array of services for applications and research involving coastal environmental prediction (Bogden et al. 2007). At the heart of the program is a service-oriented cyberinfrastructure, which includes components for data discovery, archiving, integration, translation and transport, model coupling, event notification, and resource brokering (Allen et al. 2008). Furthermore, the infrastructure developed through SCOOP is intended to support future oceanographic research by an expanded community of partners (Bogden et al. 2005). Thus, the use of standards to enable interoperability

H. Conover (✉)
Information Technology and Systems Center, The University of Alabama in Huntsville, AL, USA
e-mail: HConover@itsc.uah.edu

L. Di, H.K. Ramapriyan (eds.), *Standard-Based Data and Information Systems for Earth Observation*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-540-88264-0_6,

is key to SCOOP's success. This distributed data and information system is providing some of the standards-based information technology (IT) glue that will unite the coastal science community. Standards activities range from internal coordination among SCOOP partners to participation in national standards efforts.

6.2 SCOOP Science Motivation

The SCOOP community currently engages in distributed coastal modeling across the southeastern US, including both the Atlantic and Gulf of Mexico coasts. SCOOP partner institutions run a series of linked wind, wave, and storm surge coastal forecast models, on both a routine and event-driven basis. These model results are transmitted among the partner institutions to initialize subsequent models, stored at distributed archives and registered in the central metadata catalog (MacLaren et al. 2005).

6.2.1 Coastal Ocean Models

Figure 6.1 shows the numerical modeling grids used by the various models supported by the SCOOP infrastructure. The collection includes regular grids associated with surface-wave models – Simulating Waves Nearshore (SWAN), WAVEWATCH III (WW3), and Wave Analysis Model (WAM) – plus finite-element and curvilinear grids associated with water level or storm-surge models – Advanced Circulation Model (ADCIRC), Eulerian-Lagrangian CIRCulation (ELCIRC), and Curvilinear-grid Hydrodynamics 3D (CH3D). In general, a wave height or storm surge model or model suite is initialized for each member of an ensemble of wind forecasts (Fig. 6.2, left). For an N-element ensemble, there are N numerical forecasts run as N independent jobs on any of a variety of supercomputers available in a "Grid-enabled" network such as SURAgrid. Once the jobs are run, output from each job is transported to a distributed archive for analysis and visualization. Results are then made available for visualization and dissemination by a variety of web-accessible service and portal interfaces.

6.2.2 SCOOP Use Cases

The SCOOP architecture is designed to support an application set that includes three distinct use-case scenarios. An *event-driven, ensemble-prediction scenario* uses warnings of extreme events, such as those issued by the National Hurricane Center (NHC) during an active hurricane, to trigger automated, on-demand, ensemble-model calculations of wave height and storm surge. These calculations are designed to run quickly so they can support the urgent and immediate needs of real-time hazard planning and response. A *routine forecast scenario* runs several times daily to

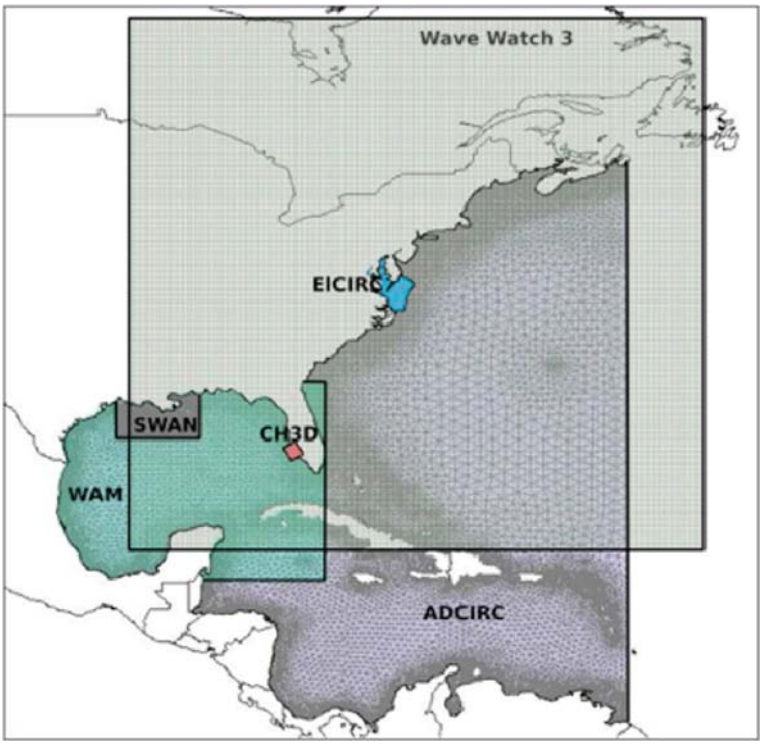

Fig. 6.1 Spatial domains and model grids used in SCOOP (figure from Bogden et al. 2007)

provide continually updated day-to-day forecasts of wave height and water level. This scenario could serve a variety of practical purposes. In SCOOP applications, the routine forecast scenario provides initialization and boundary conditions for other models in the infrastructure. A third *retrospective scenario* supports research and analysis of past events. All scenarios start with data sources that "drive" the system, including NHC warnings, wind-field predictions from operational service providers, and in situ sensor observations of variables such as water level and wave height. The coastal forecasts generated by these use-case scenarios can be visualized on a Web browser and disseminated in data formats that support decision-support tools (Bogden et al. 2007).

6.2.3 A Specific Example

The event-driven user scenario is triggered by an automated warning from the NHC. An ensemble of wind forecasts is created at one SCOOP site, and transmitted to the various coastal modeling sites. There, translation filters are used to assure the coastal models can ingest the wind data, and the coastal model runs are launched across the

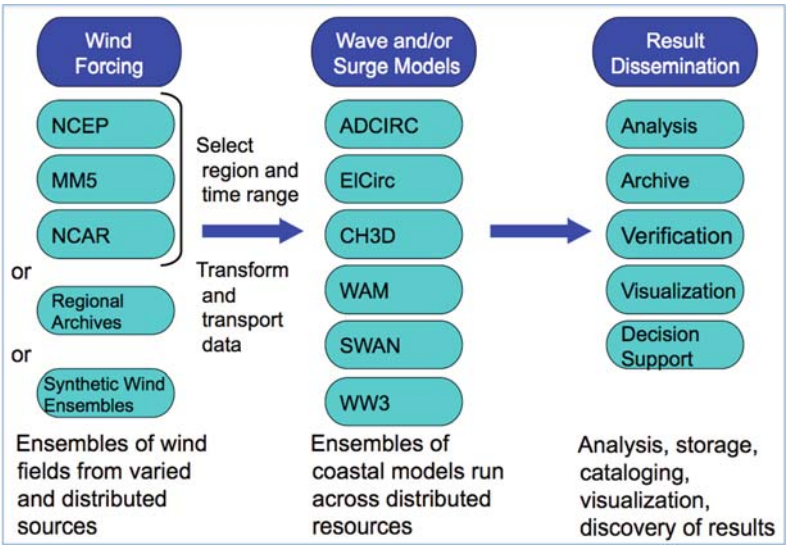

Fig. 6.2 Ensembles of atmospheric and coastal ocean models for analysis and decision support (figure from Bogden and Graves 2005)

network of distributed computing resources. Meanwhile, observation data (including water levels, waves, stream flow, and winds) are obtained from the appropriate source locations. As model results are generated, they are processed with various verification and analysis tools. Both observations and model results are pushed to the archives, cataloged, and made available to visualization services so that they can be displayed on web portals (Fig. 6.2, right) or integrated into a geographic information system (Bogden et al. 2007).

6.3 Service-Oriented Architecture

A Service Oriented Architecture (SOA) packages together independent services that have well defined interfaces. These interfaces are designed such that they are interchangeably invoked by applications or other services implemented on heterogeneous platforms and languages. Some benefits of developing applications within an SOA include re-use of existing software, reduced deployment time, minimal code changes to reduce chance of introducing errors, and orchestration of data and computational resources at distributed locations. Such an architecture allows different groups that focus on science data processing, basic research, and the applications arena to chain their various services together with others and with community-based toolkits in many different ways.

SOAs are typically built around the web services paradigm of software design and deployment. Web services are well-defined, self-contained functions that can be invoked via the Internet. The web services paradigm creates interoperable

computing assets by providing a standards-based way to describe them. Thus, web services technology provides a standard way to remotely interface with programmatic components and to orchestrate the chaining of services through standardized service descriptions.

The two main styles of web services are Simple Object Access Protocol (SOAP) and Representational State Transfer (REST). SOAP provides a standard message protocol for communication based on XML. SOAP web services have two main conventions: any non-binary attachment messages must be carried by SOAP and the service must be described using Web Service Description Language (WSDL). REST can be used loosely to describe any simple interface that transmits domain-specific data over HTTP without an additional messaging layer such as SOAP. The SCOOP SOA uses both flavors of web services. In general, SOAP services are used for communication between database and processing components, and REST services are used in web browser-based user applications, e.g., for visualization.

The SCOOP Service Oriented Architecture (SOA) is composed of a collection of modular components, each providing a well-defined functionality and communicating with the other components across standardized interfaces (Bogden et al. 2006). These components include:

- coastal models that predict phenomena such as storm surge, wind-driven waves, and inundation;
- visualization tools that facilitate efficient analysis of products;
- user friendly interfaces;
- data access, management and catalog services for input and output of data or model comparisons;
- translation and transport services that assure compatibility between the various data flows;
- computing resources that can be organized for quick turnaround of large jobs; and
- active archives of current and historical data and model results for storage, documentation, and retrieval.

The SCOOP architecture is compatible with the Global Earth Observation System of Systems (GEOSS) architecture. GEOSS, which is also based on SOA, provides a service registry where SCOOP standardized services will be published and thus made discoverable to the GEOSS community as a whole.

6.4 Use of Standards in SCOOP

The SCOOP program is committed to utilizing data and information standards wherever feasible to support system features and capabilities. A critical and overarching design principle for the SCOOP cyberinfrastructure involves adopting and implementing open community standards. This can facilitate effective partnerships among the academic, governmental and private sectors, and the end-user community.

6.4.1 Data Standards and Conventions

In order to facilitate interchange of data among SCOOP ocean models, for example to support multiple sources of input data for various ocean models, SCOOP partners determined a standard data format was needed. The *GRIB* (GRIdded Binary) format was selected as the primary format for the atmospheric forcing data (winds) used as input to the ocean models, since it was already widely used within the ocean modeling community (e.g., wind forecasts from NOAA's National Centers for Environmental Prediction – NCEP). GRIB is maintained as a standard by the World Meteorological Organization's Commission for Basic Systems (Stackpole 1994). Unidata's *netCDF* (network Common Data Form) was selected for coastal model outputs because of its increasing popularity as a community standard. NetCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data (Rew et al. 2008).

A *filename convention* was also chosen for files produced in the SCOOP system, allowing some file-level metadata – such as data type (model or observation), temporal extent, and model initialization time – to be readily accessible to users (Smith 2006). Some SCOOP applications make use of the filename convention to reduce the need for catalog queries, thus decreasing processing time.

The interdisciplinary nature of the data and the need for visualization, catalog, and other services, led to a requirement for a standard metadata convention to be used in both the metadata catalog and within generated ocean model data files. In the netCDF community, the existing *Climate and Forecasting (CF)* metadata conventions (Eaton et al. 2006) satisfy SCOOP's needs since they were designed with atmosphere and ocean forecasters in mind. These conventions extend the Cooperative Ocean/Atmosphere Research Data Service (COARDS) conventions for netCDF (COARDS 1995). CF keywords are used to label science variables within a SCOOP coastal forecast file and for geophysical parameter keywords in the SCOOP metadata catalog.

6.4.2 SCOOP Metadata Catalog and Associated Standards

The SCOOP metadata catalog is a comprehensive information repository for the project, containing high-level information on coastal observations, forecast data and the models used to generate them, data transport protocols, and other corporate knowledge. In addition, the catalog includes an inventory of all data files accessible from the SCOOP archives. The SCOOP catalog is built on a relational database schema designed to support federal metadata standards while providing for additional information needed to manage environmental observations and forecast model data. This catalog provides the metadata foundation for the various SCOOP components from data transport and archiving to model coupling, and from data discovery via search tools to visualization and data access. For example, the tropical storm component's metadata provides information about Atlantic and eastern

Pacific storms and hurricanes, including spatial and temporal information and status. SCOOP also maintains general model information to document the coastal models used. However since all scenarios start with data sources that drive the system, that information is maintained primarily in the data collection metadata utilizing the Federal Geographic Data Committee (FGDC)'s Content Standard for Digital Geospatial Metadata (FGDC 1998) and the CF vocabulary.

Data providers are asked to supply data collection metadata via a registration form that populates the underlying database tables associated with data collection, online access and services, and data provider information. This data collection information conforms to several aspects of the FGDC standard: (1) Identification Information, (2) Spatial Reference Information, and (3) Metadata Reference Information. The data provider (4) Contact Information is linked with any associated data collections. The Content Standard for Digital Geospatial Metadata Workbook (FGDC 2000) is an excellent requirements guide for specifying the metadata that should be recorded for each data collection.

Basic *Identification Information* includes the data collection name, collection citation and description, collection content time period, spatial domain, status, access, and use constraints. The *Spatial Reference Information* captures the geographic information for all collections using the geographic coordinate units and the latitude and longitude resolutions. The in situ observational data is additionally characterized by use of the FGDC vertical coordinate system information specifying the altitude or depth and its datum name, e.g., "NAVD88" for altitude. *Metadata Reference Information* required by the FGDC captures the date the metadata became available, the metadata contact, and the metadata standard that is used to describe the data. SCOOP extends its documentation of the data collections further by maintaining a list of vocabularies used for the physical parameters, location names and other specific attributes not defined by FGDC but used in the SCOOP metadata catalog. For example, for the "location" attribute, the Global Change Master Directory "Location Keywords" (Olsen et al. 2007) were used; for geophysical parameters, "COARDS/CF Standard Names" were referenced. The use of these vocabularies aids in consistency of terminology throughout the SCOOP program. FGDC requires data provider information for data collections. The SCOOP metadata schema stores this information according to the FGDC *Contact Information* attributes, which include name, position, and contact information.

While FGDC specifies the types of metadata attributes to be cataloged for each data collection, SCOOP is also relying on more detailed standards for how to specify some of these attributes. For example, the FGDC standard allows attributes designating time periods in free text, but SCOOP chose to use the ISO 8601 standard for the representation of dates and times to maintain consistency throughout its catalog. SCOOP maintains several time fields using the ISO 8601 duration notation, including file update frequency, the length of time the data covers within the file (file time period), and the time between individual data updates within the data collection (data temporal resolution). For example, the ADCIRC model outputs data hourly and thus has temporal resolution designated as "PT1H" (period of time 1 h); the ELCIRC model's data file contains 5 days of data ("P5D" – period of five days), and NCEP wind data files are updated every 6 h ("PT6H").

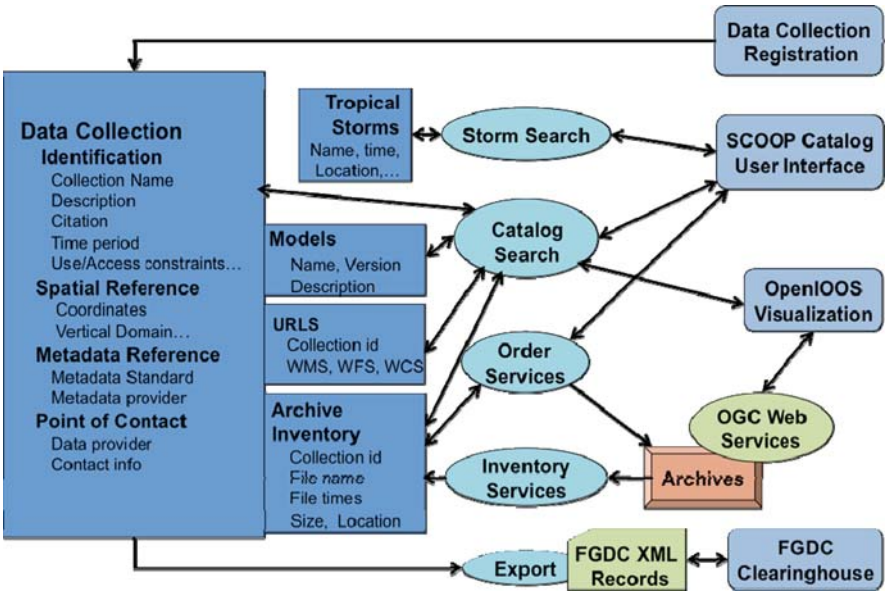

Fig. 6.3 SCOOP information system

Figure 6.3 is a high-level graphical representation of the SCOOP information system, with rectangles representing database tables, ovals representing catalog services, and rounded rectangles representing user interfaces to the system. The catalog search application (http://scoop.sura.org/Catalog/) allows the user to search for any SCOOP data collection by keyword, spatial or temporal attributes. From the resulting list of data collections, the user can access the collection description including FGDC metadata, list or view data files, and order the data from the archives. Another catalog service exports information on SCOOP data collections in a schema compliant with the FGDC metadata standard. SCOOP hosts an FGDC Clearinghouse node, which provides the exported SCOOP metadata to the public through Geospatial One-Stop and via searches using Z39.50, an ANSI/NISO standard protocol for information retrieval (ANSI 1992).

6.4.3 Web Services for Catalog Search

The catalog search pages and other applications access the SCOOP metadata database via a suite of web services. SCOOP catalog services also provide for automated interactions by other SCOOP partners such as the archives. Supported web services standards include WSDL, SOAP and XML, all maintained by the World Wide Web Consortium (http://www.w3.org/). Catalog web services are written in either Java or Perl, and each is tested locally with both Java and Perl clients to assure compliance to the SOAP and WSDL standards. The XML message schemas

for these services also incorporate selected standards; for example, a subset of ISO 8601, xsd:dateTime, is used for fully qualified date/times including time zone information.

The inventory services are designed in conformance to the Basic Profile specifications (Ballinger et al. 2004a, b) of the Web Services Interoperability Organization (http://ws-i.org/), an open industry organization chartered to promote web services interoperability across platforms, operating systems, and programming languages. Thus, the service interfaces provide maximum interoperability. Since the service request and response parameters follow a specified XML schema, use of the "Document" style for SOAP services is a natural choice (Butek 2005). The Document style generally uses literal encoding, meaning that the SOAP Body message contents conform to a specific XML Schema. Furthermore, Document style services are known to have better performance than other styles. The WSDLs generated are simple and easy to use in creating the SOAP Body message. For each of the services' available methods, test clients are provided in Java and Perl/SOAP::Lite using the public WSDLs.

SCOOP plans to provide SCOOP metadata through Catalogue Services for the Web (CSW), an OGC standard protocol for data and service catalogs (Nebert et al. 2007). The *Open Geospatial Consortium (OGC)* comprises international commercial companies, government agencies, and universities, using a voluntary consensus process to collaboratively develop open standards for geospatial data and information services. SCOOP's initial investigation is focused on available reference implementations of CSW from *GeoNetwork* (http://geonetwork-opensource.org/) and *deegree* (http://www.deegree.org/). Both of these reference implementations use the ISO 19115/ISO 19139 metadata profile for CSW, also adopted by the OGC, as one of their base application profiles. ISO 19115 defines a schema for describing geographic information and services, while 19139 provides a concrete XML implementation specification for 19115. Eventually, the CSW implementation of the SCOOP Catalog will be registered with GEOSS and discoverable through their portal.

6.4.4 Web Services for Data Access and Visualization

An important capability in SCOOP is the access to and display of geospatial information, in particular coastal model results. SCOOP is working to enable visualization of all the archived data products, including forecasts generated by SCOOP participants, forecasts from external organizations such as NCEP winds, and coastal observational data. Within SCOOP, these products are catalogued and archived to support their use by decision support systems and by other models. To make this information as universally accessible as possible, SCOOP has adopted standardized service interfaces for access to data stored in different data formats, using different data models, coordinate reference systems, geometry models, etc. The OGC has guided the geospatial community in defining and standardizing service interfaces for

access to data, images, features, etc. These interfaces have gained wide acceptance as standards for open access to geospatial data. Two of the OGC service interfaces that have been successfully implemented and utilized in the SCOOP architecture include the *Web Map Service (WMS)* and *Web Feature Service (WFS)* interfaces.

The WMS specification (de la Beaujardiere 2006) provides a simple HTTP interface for requesting georegistered map images from one or more distributed geospatial data sources, with a response providing georegistered map images that are compatible with conventional browser applications. OGC-compliant client applications are able to use WMS requests to retrieve map images that can be layered with other image requests to compose more complicated map presentations. The SCOOP project has been successful in utilizing WMS to support its core visualization for coastal model forecasts. By providing WMS for these data, SCOOP is able to support the display of this information in OGC-compliant visualization applications, such as the OpenIOOS.org application, that demonstrate the integration of OGC services for visualization of geospatial model data, shown in Fig. 6.4.

Similarly, the WFS specification (Vretanos 2005) supports the request and retrieval of geospatial features and associated metadata. While WMS is useful for

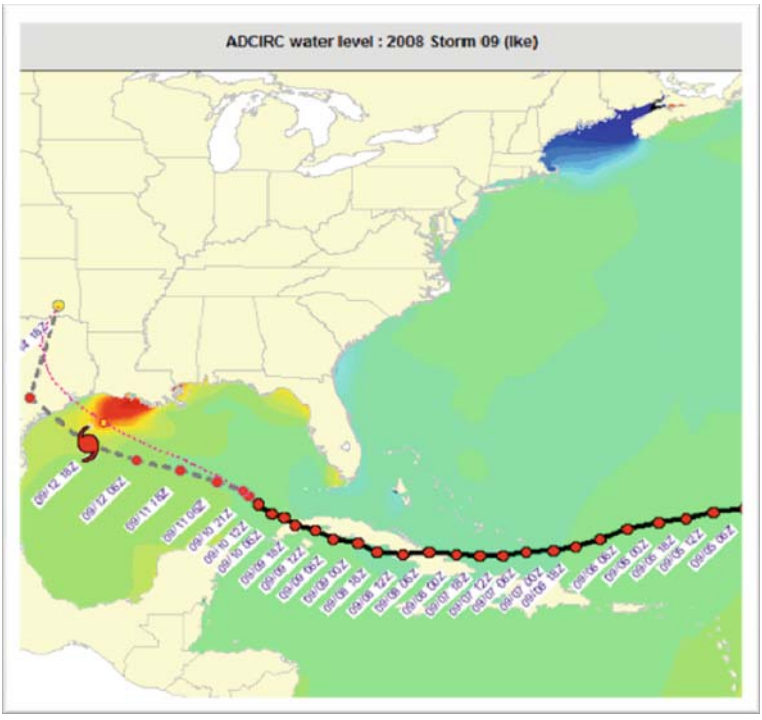

Fig. 6.4 Example of water level forecast from ADCIRC model (via WMS) overlaid with storm track for Hurricane Ike in 2008 (via WFS)

displaying images representing data such as the output of SCOOP coastal models, WFS can provide vector information (e.g., points for observation sensors) coupled with data for each point. WFS provides the ability to generate interactive displays allowing users to request additional information about specific features on a map. For example, each of the points in the storm track shown on the map in Fig. 6.4 represents information such as date, time, and storm location.

WMS support for many of the SCOOP coastal models has been implemented at the SCOOP archive locations where the data is located. The close proximity of the service to the data avoids the overhead of data network transfer, thus allowing the response time for the services to be faster in support of visualization client applications. While WMS output has been useful for client applications, the request specification is somewhat lacking in supporting requests for model data. For coastal modeling specifically, routine model runs typically are scheduled for every 6 h and produce predictions for some extended period in the future (e.g., 36 h or 72 h). Thus, forecasts for coastal conditions for a given date and time will be generated by several different model runs over the preceding days. Applications requesting visualization of this data need to be able to request not only the time range of data to be displayed, but also the time of the forecast cycle (which model run). Since the current WMS specification only allows for the begin/end time range parameters, the SCOOP project has adopted application-specific extensions to the WMS specification to support the finer definition of time necessary to distinguish between model runs for overlapping time periods. Requests using the default WMS specification are supported and return valid results, but client applications do not have full control over time specification unless the SCOOP extensions are used.

The SCOOP metadata catalog maintains WMS and WFS service URLs as part of the metadata for SCOOP data collections. While not a true service registry, the data catalog does provide the ability to search and discover the services associated with specific SCOOP collections, if defined by the data provider. These discoverable services serve as the basis for search applications designed for human interaction. For example, a hurricane-centered data search tool, requested by SCOOP coastal modeling partners, allows users to specify a region of interest by storm name and year, instead of or in addition to specifying geospatial and temporal coordinates. Storm information is provided via a WFS. The resulting storm track is displayed on an interactive map where a user may view attributes such as wind speed, pressure and storm category (Fig. 6.5). The user may also select specific spatial and temporal subsets of the storm to help refine searches against the data catalog for information on forecasts and observations of the displayed storm.

Verification of model output is an important step in trying to improve the quality of predictions. In the case of coastal models, SCOOP provides capabilities to support model-to-observation and model-to-model comparisons, with the observations being readings from coastal buoys for water level and wave height, speed and direction. To facilitate these comparisons, the SCOOP archives have implemented filters that extract model values corresponding to the known buoy locations from the model output. These extracted points are then made available to applications providing the model-to-observation comparisons (e.g., OpenIOOS.org) through a WFS interface

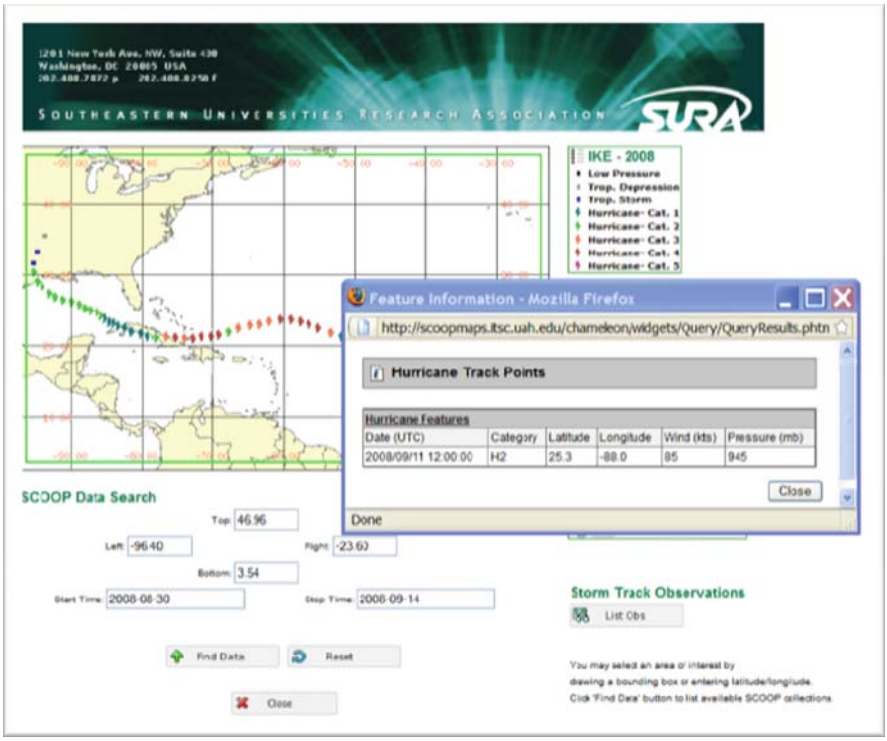

Fig. 6.5 Storm-based search display of Hurricane Kyle (2008) using WFS to access tropical storm database

(see Fig. 6.6). WFS is appropriate in this case as the extracted model values are represented as geospatial point features with the associated value as attribute information. The graphical representation of the WFS information in this case is in the form of a line graph comparing the model points to the corresponding observation points for the selected buoy. The buoy information for model-to-observation comparisons is derived from a number of sources, and where supported is also accessed via WFS, but this is not available in all cases.

6.5 SCOOP Participation in Standards Activities

SCOOP partners are key participants in a number of different ocean science community standards activities. These activities range from community review of proposed data system standards to experimentation with promising protocols and technologies.

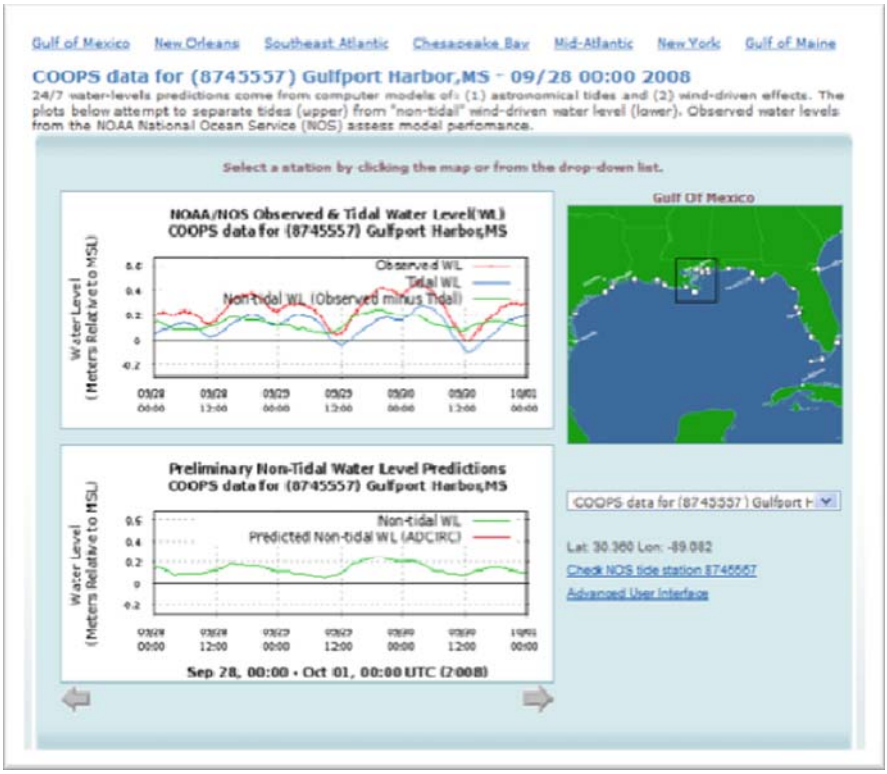

Fig. 6.6 Model-to-observation comparison chart from OpenIOOS.org

6.5.1 IOOS Data Management and Communications

The Integrated Ocean Observing System (IOOS) began almost two decades ago as a multi-agency initiative to implement a system of systems that will routinely and continuously provide quality controlled data and information on the oceans and Great Lakes from the global scale of ocean basins to local scales of coastal ecosystems. One of the highest priorities for the evolving IOOS is a Data Management and Communications (DMAC) subsystem, which has been advanced over the last decade by the multi-agency and multi-sector US IOOS DMAC Steering Team. The DMAC subsystem is intended to knit together the global and coastal components of the IOOS, and link every part of the observing system from the instruments to the users. The DMAC subsystem will transmit multidisciplinary observations collected from a broad range of sensors and platforms to diverse user communities (Hankin and Steering Committee 2005). SCOOP partners are participating along with other IOOS stakeholders in DMAC planning and assessment activities to ensure that current and future community needs and priorities are addressed.

Currently, a primary activity for DMAC is implementation of a standards process (DMAC 2006) to identify and review technologies and protocols in use within current ocean sciences data systems, in order to support data and systems interoperability within the IOOS DMAC infrastructure. DMAC recommends standards and practices in a variety of areas including metadata and data discovery, data transport, and data archive and access. SCOOP has provided a member to the DMAC Metadata and Data Discovery Expert Team. This team is charged with determining IOOS requirements for metadata, reviewing and providing recommendations on potential IOOS metadata standards, and helping to develop crosswalks between different metadata standards. SCOOP's participation on the DMAC Metadata Expert Team will bring SCOOP experience and lessons learned to the wider ocean research community, as well as assure SCOOP's compliance with emerging ocean metadata standards.

6.5.2 Marine Metadata Interoperability Project

In 2004, NSF funded the Marine Metadata Interoperability project (MMI), a community effort whose goal is to promote agreements, standards, and best practices for sharing data among the marine community (Graybeal et al. 2005). SCOOP partners are active participants in MMI, helping to identify best practices to make science data easy to distribute, advertise, reuse, and combine with other data sets. A primary goal of MMI is to develop *ontologies* in order to bridge the different vocabularies already in use by the ocean sciences community. MMI is providing a robust semantic mediation framework for oceans, including a semantic mediation architecture, guidance for constructing, mapping, and retrieving vocabularies, and associated tools. One such tool is a vocabulary registry, based on semantic web technologies, to register, map, and access and visualize ontologies.

6.5.3 OpenIOOS

SCOOP is a major participant in the OpenIOOS Interoperability Demonstration, one of a number of coastal sciences community efforts, with the goal of integrating national and regional activities into a seamless tapestry of observations and predictions. Partners include several federal agencies and dozens of the top research universities in the country. OpenIOOS relies heavily on OGC standards such as the WMS and WFS protocols to display near real time coastal observations together with water level, wave, and surge forecasts (e.g., Figs. 6.4 and 6.6). The OpenIOOS initiative began in 2003, when an informal community of ocean-observing program participants gathered to investigate the capabilities of the WMS and WFS specifications. They demonstrated interoperability across institutions with web-mapping products that included in situ and satellite measurement of sea-surface temperature. Their shared website, http://www.openioos.org, has been providing these real-time

sea-surface temperature maps since early 2004, using readily available software. OpenIOOS has served a both a community building tool and a technology demonstration for the ocean sciences community. SCOOP has leveraged the OpenIOOS web site as its primary visualization user interface, and has, in turn, contributed coastal forecasts and other data to the OpenIOOS system.

6.5.4 OOSTethys

While SCOOP is primarily focused on coastal modeling, the program also relies heavily on ocean observations. The ocean-observing community involves scientists from academia, industry, and government programs that individually deploy a wide variety of sensors. This variety introduces obstacles to creating seamless and coordinated access to observations from these disparate and heterogeneous sources. Simplified data exchange can improve the way scientists observe the oceans and inform their management. Additionally, data sharing challenges become more difficult if they are to respect the need for experimental reproducibility – a hallmark of the scientific process – because different groups often represent, transport, store and distribute their data in different ways.

In order to address these data sharing issues, MMI and SCOOP participants agreed to combine efforts in implementing and demonstrating data access protocols, initiating the OOSTethys project in 2006. The OOSTethys initiative has since grown well beyond the original partners to include international involvement. OOSTethys has evolved as an open source collaborative project that seeks to create and adapt tools to help developing observing system components (Bermudez et al. 2006). The primary goal of OOSTethys is to dramatically reduce the time data service providers spend installing, adopting, and staying abreast of new technologies and standard services. OOSTethys has three related thrust areas, each one supporting the others:

1. *Open-source standards-compliant software*: These are reference implementations that allow data providers (e.g., Regional Associations, NOAA programs, other state or federal agency programs) to serve data with web services that comply with a common set of consensus standards and best practices. OOSTethys is focusing on OGC standards for realtime observed data. It has provided Java, Perl and Python toolkits, and has helped advance specifications from the OGC Sensor Web Enablement initiative.
2. *OpenIOOS Test Bed*: A collection of services needed for creating end-to-end systems of systems, that facilitates developing and testing of reference implementations. SCOOP contributions include a service registry, metadata catalog, data aggregator, and a portal. The portal, OpenIOOS, is meant to be an operational client, visualizer and decision support tool. MMI provides a semantic mediator component.
3. *Open community process*: OOSTethys encourages broad participation, standards compliance, and minimal duplication of effort. OOSTethys uses collaborative

open tools to moderate decision making, prioritize tasks and engage the community. OOSTethys members participate in other projects to advance standards (e.g., OGC working groups, GEOSS Architecture Implementation Pilots). OOSTethys members also initiate activities such as OGC Ocean Science Interoperability Experiment.

6.5.5 OGC Oceans Interoperability Experiment

Several of the OGC members involved in OOSTethys (SURA, Monterey Bay Aquarium Research Institute, GoMOOS, Texas A&M and Unidata, an NSF-funded program for sharing Earth science data and tools) initiated the companion OGC Ocean Science Interoperability Experiment (Oceans IE) in 2007. The goal of this experiment is to leverage the OGC Interoperability Program for advancing standards-compliant best practices for publishing of marine observations. Participation in the OOSTethys initiative has grown because of this formal connection to the OGC and its relatively large international and multi-sector membership. The timing of both OOSTethys and Oceans IE coincided with advances in the OGC Sensor Web Enablement (SWE) initiative, which provides standards for data and information acquisition from sensor systems and data repositories. The OGC SWE standards framework provides specifications for interfaces, protocols and encodings that are designed to enable implementation of interoperable, service-oriented networks of sensors and applications (Botts et al. 2007). Providing such standard interfaces to sensor data can minimize the custom software required for management, visualization and analysis of different types of sensors and observations.

The availability of the new SWE protocols prompted a comparison between the WFS standard and the relatively new SWE Sensor Observation Service (Na and Priest 2007). The Oceans IE Phase I investigated the use of WFS and SOS for representing and exchanging point data records from fixed in situ marine platforms (Bermudez 2008). It concluded that SOS was better suited than WFS for this purpose. By publishing an SOS service instead of a WFS service, communities will not be required to create and maintain message schemas, and interoperability at the client side is achieved; however, this requires an effort in creating and maintaining controlled vocabularies by marine communities. Extensive investigation, software development, and real-world testing resulted in the set of open source SOS reference implementations and community cookbooks on http://www.oostethys.org/. The Oceans IE also developed the following best practices for using an OGC Sensor Observation Service (v1.1), which will help improve existing standards and recommendations at OGC:

- Requesting “latest observation”;
- Encoding of OGC Uniform Resource Names (URNs) when versioning is missing;
- Publishing of Uniform Resource Identifiers (URIs) by service providers;

- Using Semantic Web technologies to categorize SOS services;
- Publishing an SOS as an "HTTP-Get" service;
- Encoding vertical datums (sea level based systems, geoid based systems and bottom based systems) in marine observations.

Through SURA, OOSTethys and OpenIOOS, several SCOOP partners are key participants in Oceans IE. While SCOOP's point data servers (e.g., for buoy observations or storm tracks) and visualization clients are already implemented to use the WFS protocol, this recent experimentation with SOS implementations can lead to wider interoperability with others.

As the standards and experience of the different groups on publishing and accessing the SWE standards evolve, the OOSTethys tools will continue to improve. The importance of OOSTethys and the Oceans IE can be measured by the adoption of these technologies by important ocean observing systems initiatives. SOS is being considered or adopted, at the time of writing this document, by NSF's Ocean Observing Initiative, the multi-agency and multi-sector US IOOS DMAC Steering Team, the NOAA/IOOS Data Integration Framework, and Europe's ESONet program.

6.6 Conclusions

Standard interfaces, including common file formats, metadata schemas and vocabularies, catalog web services, and OGC data access and web visualization services, support development of a variety of specialized SCOOP applications and user interfaces. Close cooperation between the information technology and coastal science modeling communities is producing positive results toward a real-time modeling environment that will benefit coastal stakeholders through better predictive capabilities.

Acknowledgments This work is part of the Southeastern Universities Research Association Coastal Ocean Observing and Prediction (SCOOP) Program funded by the Office of Naval Research Award N00014-04-1-0721, NOAA Ocean Service Award NA04NOS4730254, and National Science Foundation Grant 0607431. The authors would like to acknowledge their SCOOP colleagues, especially those partners who have collaborated in testing and deploying the standards and web services described here within SCOOP's service oriented information architecture.

References

Allen G, Bogden P, Kosar T, Kulshrestha A, Namala G, Tummala S, Seidel E (2008) Cyberinfrastructure for coastal hazard prediction. In: CTWatch Quarterly, 4(1)

American National Sandards Institute (1992) ANSI/NISO Z39.50-1992: American National Standard Information Retrieval Application Service Definition and Protocol Specification for Open Systems Interconnection. NISO Press, Bethesda, MD

Ballinger K, Ehnebuske D, Ferris C, Gudgin M, Liu C, Nottingham M, Yendluri P, eds (2004a) WS-I basic profile 1.1

Ballinger K, Ehnebuske D, Gudgin M, Nottingham M, Yendluri P, eds, (2004b) WS-I basic profile 1.0

Bermudez L, ed (2008) OGC Ocean science interoperability experiment phase 1 report, OGC Engineering Report 08-124

Bermudez L, Bogden, P, Bridger P, Forrest D, Graybeal J, Creager G (2006). Toward an ocean observing system of systems. In: Proceedings of the Oceans'06 MTS/IEEE, Boston, MA

Bogden P, Graves S, (2005) SCOOP an IOOS testbed. Presented: 2005 Fall Meeting of the American Geophysical Union, San Francisco, CA

Bogden P, Allen G, Stone G, Bintz J, Graber H, Graves S, Luettich R, Reed D, Sheng P, Wang H, Zhao W (2005) The Southeastern University Research Association Coastal Ocean observing and prediction program: integrating marine science and information technology. In: Proceedings of the Oceans 2005 MTS/IEEE Conference, Washington, DC

Bogden P, Conover H, Creager G, Flournoy L, Allen G, Blanton B, Graber H, Graves S, Luettich R, Stone G, Perrie W, Sheng YP, Wang H, Zhao W (2006) The SCOOP service-oriented architecture for ocean observing and prediction. In: Proceedings of the Oceans'06 MTS/IEEE, Boston, MA

Bogden P, Gale T, Allen G, MacLaren J, Almes G, Creager G, Bintz J, Wright L, Graber H, Williams N, Graves S, Conover H, Galluppi K, Luettich R, Perrie W, Toulany B, Sheng YP, Davis J, Wang H, Forrest D (2007) Architecture of a community infrastructure for predicting and analyzing coastal inundation.*Marine Technology Society Journal*, 41(1): 53–71

Botts M, Percival G, Reed C, Davidson J (2007) OGC sensor web enablement: overview and high level architecture. OGC White Paper 07-165

Butek R (2005) Which style of WSDL should I use? Online: http://www-128.ibm.com/developerworks/webservices/library/ws-whichwsdl/

Cooperative Ocean/Atmosphere Research Data Service (1995) Conventions for the standardization of NetCDF files. Online: http://ferret.wrc.noaa.gov/noaa_coop/coop_cdf_profile.html

de la Beaujardiere J, ed (2006) OpenGIS® Web Map Service (WMS) implementation specification

DMAC Steering Committee (2006) IOOS-DMAC guidelines/standards adoption process. Online: http://dmac.ocean.us/dacsc/docs/may_2006/IOOS_DMAC_Standards_Process_053106a.doc

Eaton B, Gregory J, Drach B, Taylor K, Hankin S (2006) NetCDF climate and forecast (CF) metadata convention. Online: http://www.cgd.ucar.edu/cms/eaton/cf-metadata/

Federal Geographic Data Committee (1998) Content standard for digital geospatial metadata. FGDC-STD-001-1998

Federal Geographic Data Committee (2000) Content Standard for Digital Geospatial Metadata Workbook, Version2.0

Graybeal J, Bermudez L, et al. (2005) Marine metadata interoperability project: leading to collaboration. Presented: Local to Global Data Interoperability – Challenges and Technologies Symposium, Sardinia, Italy

Hankin S, Steering Committee (2005) Data management and communications plan for research and operational integrated ocean observing systems: I. interoperable data discovery, access, and archive. Online: http://dmac.ocean.us/dacsc/imp_plan.jsp

MacLaren J, Allen J, Dekate C, Huang D, Hutanu A, Zhang C (2005) Shelter from the storm: building a safe archive in a hostile world. In: Proceedings of the Second International Workshop on Grid Computing and its Application to Data Analysis (GADA'05), Agia Napa, Cyprus, Springer Verlag

Na A, Priest M, eds (2007) Sensor observation service. OpenGIS® Implementation Standard OGC 06-009r6

Nebert D, Whiteside A, Vretanos P, eds (2007) OpenGIS® Catalogue Services Specification, version 2.0.2. OGC 07-006r1

Olsen LM, Major G, Shein K, Scialdone J, Vogel R, Leicester S, Weir H, Ritz S, Stevens T, Meaux M, Solomon C, Bilodeau R, Holland M, Northcutt T, Restrepo RA (2007) NASA/Global Change Master Directory (GCMD) Earth science keywords, version 6.0.0.0.0

Rew R, Davis G, Emmerson S, Davies H, Hartnett E (2008) NetCDF users guide. University Corporation for Atmospheric Research
Stackpole J (1994) Guide to WMO binary code from GRIB 1", Technical Report No. 17, Geneva
Smith M (2006) SCOOP filename conventions. Online: http://scoop.sura.org/documents/naming_convention_final_5-3-06.pdf
Vretanos P, ed (2005) OpenGIS® Web Feature Service (WFS) Implementation Specification

Chapter 7
A New Approach to Preservation Metadata for Scientific Data – A Real World Example

Ruth Duerr, Ron Weaver, and Mark A. Parsons

7.1 Introduction

The Open Archival Information System (OAIS) Reference Model was developed by the Consultative Committee for Space Data Systems (CCSDS) in the late 1990s and was adopted as an ISO standard in 2003 (ISO14721:2003) (CCSDS 650.0-B-1, 2002). Recently, many libraries, data centers, and archives around the world have started to adopt this archive model. As a notable example, the National Oceanographic and Atmospheric Administration (NOAA) of the United States of America adopted the model for their Comprehensive Large Array-data Stewardship System (CLASS). CLASS is expected to be the primary repository and access portal for NOAA's Earth science, satellite-remote-sensing data (http://www.class.noaa.gov/nsaa/products/welcome). Because of the huge scale of these efforts, it is important to carefully consider the manner in which data archives implement the OAIS Reference Model. The specifics of the implementation could impact the preservation and therefore the use and usability of vast quantities of data for many years.

The National Snow and Ice Data Center (NSIDC) undertook the development of an prototype operations and preservation metadata tool based on the OAIS Reference Model and compatible with the PREservation Metadata Implementation Strategies (PREMIS) Data Dictionary (PREMIS Working Group, 2005) in order to consolidate the operations and preservation metadata collected for many of NSIDC's data. The results from this prototype include an assessment of the applicability of the PREMIS standard in the context of a science data archive, a set of proposed extensions to the existing NSIDC metadata database, detailed metadata development and maintenance procedures, and a set of lessons learned.

R. Duerr (✉)
National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Science, University of Colorado at Boulder, CO, USA
e-mail: rduerr@nsidc.org

L. Di, H.K. Ramapriyan (eds.), *Standard-Based Data and Information Systems for Earth Observation*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-540-88264-0_7,

7.2 Overview

Unlike many standards, the OAIS Reference Model is neither a design nor implementation level specification. Instead it provides a high-level set of archive requirements and functional and information models that describe the range of activities and data transformations necessary to keep data available and useful into the future, as well as definitions for archive-related terminology. Furthermore, the model states that other groups should tackle the creation of the subsidiary protocols and standards necessary for detailed archive design and implementation. Clearly, the Comprehensive Large Array-data Stewardship System (CLASS) and other archive systems need to develop their own detailed implementation standard protocols, requirements, and procedures, but wise data stewardship requires that data managers be aware of other relevant standards and look to harmonize these standards where possible.

One area where development of subsidiary archive standards has occurred is in the definition of preservation metadata. The Preservation Description Information (PDI) of the Open Archival Information System (OAIS) Reference Model is a high-level set of metadata requirements necessary to ensure the long-term preservation of data. While it could potentially be argued that for Earth science data many of the PDI requirements are covered by existing detailed metadata standards such as the ISO 19115 (ISO, 2003) and Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) standards (FGDC-STD-001-1998), until recently there was no standard available that is broadly applicable to all types of digital data. To address this gap, the Online Computer Library Center (OCLC) and the Research Libraries Group, Inc. (RLG) jointly sponsored the international PREMIS Working Group (May 2005). In May 2005, the group published their final report (PREMIS Working Group, 2005), containing version 1.0 of a data dictionary for preservation metadata. Since then, XML schemata have been developed; the standard and schemata are being maintained by the Network Development and MARC Standards Office of the Library of Congress (LOC, Schemas for PREMIS); an international Editorial Board has been established to provide direction for the continued development and maintenance of the standard; and an implementation discussion group and associated registry formed. Version 1.0 of the data dictionary underwent a one-year trial period during which a number of revisions were proposed based on the results of implementation by a variety of organizations. These suggested revisions were reviewed by the Editorial Board for inclusion in the next version of the standard, which was published in early 2008 (PREMIS Working Group, 2008).

However, it is significant to note that there was little direct input from scientific data managers on the original PREMIS working group, and the majority of the participants represented the international library community. Several US agencies were involved, including participation by the National Archives and Records Administration (NARA), and a member of the Commerce, Energy, NASA, Defense Information Managers Group (CENDI) gave advice, but there was no direct participation by any of the national data centers in the United States. Moreover, the composition of the Editorial Board indicates the same range of participation today,

though it must be noted that the original working group tendered an open call for participation.

Unknowingly, the National Snow and Ice Data Center (NSIDC) was pursuing a similar preservation metadata effort in parallel with the PREMIS working group. In 2002, NSIDC recognized a need to enhance its ability to manage scientific data for the long term and adopted the OAIS Reference Model as a guide particularly in the area of preservation metadata. Thereafter, NSIDC drafted a preliminary preservation metadata schema based on the OAIS preservation metadata definitions (see Table 7.1). In the summer of 2005, NSIDC began a preservation metadata implementation case study. We endeavored to reconcile the NSIDC and PREMIS schemata and apply them to a diverse set of data sets from the Cold Land Processes Experiment (CLPx) (Cline et al., 2003). Data sets from CLPx were chosen for our case study because they are temporally and spatially constrained, but include a broad array of scientific data types from ground sampling to airborne and satellite remote sensing. The results from this case study included a set of proposed extensions to the existing NSIDC metadata database, detailed metadata development and maintenance procedures, and a set of lessons learned. More recently, these extensions to the existing NSIDC metadata database have been prototyped along with a user interface. The results of these activities are one of the first tests of the applicability of the PREMIS preservation metadata schema to science data sets and might provide guidance to other scientific data centers and systems.

In the following section, the topic of science data preservation is briefly described, the relevant metadata standards outlined along with their arrangement in the larger preservation and access landscape. An overview of NSIDC and its pre-existing metadata-related infrastructure follows. Section 7.4 describes NSIDC's activities in this area, discussing the procedures used, and detailed findings concerning the applicability of the PREMIS metadata to NSIDC's data sets.

Table 7.1 Candidate reference information metadata items

Name	Definition	PREMIS field (if any)
Aliases/ nicknames	A table of aliases, nicknames, and former titles for the data set; including descriptions of when and where the alias was used; who made the association to this data set (and when); and what material (documentation, personal reference etc.) was used in making the association	Object identifier
Data set title	Title of the data set	Object identifier
DIF ID	Identifier for the data set in the GCMD	Object identifier
FD number	Identifier of the paper file folder containing information about this data set	Object identifier
Data center affiliation	The history of which data centers and programs within the overall NSIDC umbrella contributed to the maintenance of this data set; along with accompanying description of the reason for changes over time	A series of events potentially linked to agents with permission statement changes

7.3 Background

Clearly data are at the heart of science, key to enabling results replication – a core tenet of the scientific method. However, the need to explicitly provide preservation of and access to scientific data is fairly recent, dating back to the beginning of the computer age and the ever-increasing volumes of digital data. Prior to this era, data were either simply unavailable, recorded in a scientist's files or notebooks, or published in one of the many discipline-specific journals or report series that had data publication in their charter. Back then access may have been time consuming and difficult; but, preservation for published data, arguably the most important data, was straightforward – libraries were responsible for maintaining their collections of journals and world-wide the number of libraries maintaining a particular journal could be very large. While individual copies may have disappeared due to misadventure or neglect, the shear number of copies available was sufficient to ensure the longevity of most works.

Preservation and access to digital data has never been that simple. As a report sponsored by the National Science Foundation and the Library of Congress in 2002 puts it "digital objects require constant and perpetual maintenance, and they depend on elaborate systems of hardware, software, data and information models, and standards that are upgraded or replaced every few years (NSF, August 2003)."

Given this constant technological change, it took the accumulation of several decades of experience from a wide variety of organizations responsible for managing digital objects in order to formulate the beginnings of a theoretical framework for digital preservation. This framework, encapsulated in the OAIS reference model, which is described further in the next section, provides both information and functional models for preservation of digital objects. Currently, the PREMIS metadata standard (see Sect. 7.3.3) is the only standard that attempts to fulfill the information requirements described by the OAIS reference model.

While preservation of digital objects in general is complex, the situation is even further complicated for digital science data. It isn't enough to preserve the data, a whole host of contextual data and information also needs to be preserved and made accessible in order to make the data useful both now and in the future. While this contextual information varies somewhat from discipline to discipline, for the kinds of environmental data archived at the NSIDC, the requirements for this contextual information were spelled out over a decade ago in a report from a workshop held under the auspices of the US Global Change Research Program Office (Hunolt, 1999) and are largely captured in the FGDC CSDGM metadata standard (Federal Geographic Data Committee, 1998) and its international successor the ISO 19115 family of metadata standards (ISO, 2003) (see Sect. 7.3.2).

7.3.1 OAIS Reference Model

Originating in the decades of experiences with digital data management of the major space agencies of the world, the OAIS reference model provides a conceptual framework for understanding the responsibilities of an archive along with

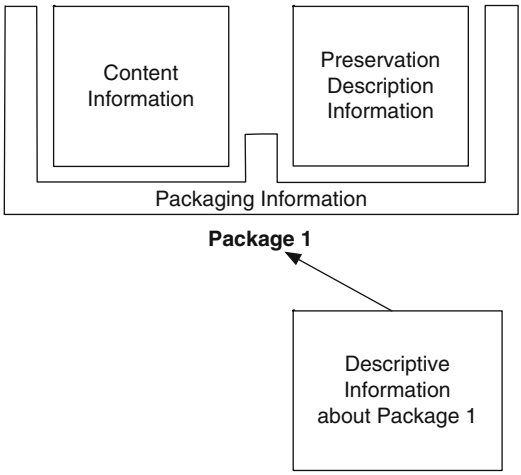

Fig. 7.1 OAIS information package concepts and relationships, CCSDS 2002

high-level descriptions of the information and functionality required in order to fulfill these responsibilities. Of these components, this work focuses solely on the OAIS information model.

At the core of the OAIS information model is the concept of an information package (see Fig. 7.1). An information package in its most basic form consists of the Content Information to be preserved along with its Preservation Description (see definitions below). While not necessarily a part of an Information Package, Packaging Information provides the physical and logical binding between the Content and its Preservation Description, while Descriptive Information allows the broad user community to find, assess, and access the Package.

- *Content Information* consists of the digital object to be preserved, along with enough representational information about the object (i.e., syntactic and semantic information) so that the designated user community can independently understand the object without resorting to expert guidance.
- *Preservation Description Information* contains information about the provenance and lineage of an object such as where it came from, how it was created, and the chain of custody since; contextual information detailing why this object was created and how this object relates to other objects and the larger universe of information; reference information defining the naming and identification systems that uniquely identify the object; and the fixity techniques used to verify that the object has not been changed in an undocumented way since its creation.
- *Packaging Information* maintaining the physical or logical association between the content information and its preservation description, describing information such as the media, directory structures, and file naming conventions used.
- *Descriptive Information* is the set of information about an object that allows a user to discover it and determine whether it is of interest. Often, much of the descriptive information can be derived from the Content Information and Preservation Description Information.

7.3.2 FGDC CSDGM and ISO 19115

The predominate metadata standard used by the Earth Science community is the "Content Standard for Digital Geospatial Metadata" established by the Federal Geographic Data Committee (FGDC) and mandated for use for all federally funded geospatial data (Federal Geographic Data Committee, 1998). The FGDC content standard and its more recently defined international equivalent, the ISO 19115 family of standards (ISO, 2003), defines the metadata needed in order for users to find, assess, access, and use geospatial data. This is essentially equivalent to the Descriptive Information in the OAIS information model. One type of information that is not explicitly targeted is that needed to ensure the long-term preservation of the data described, precisely the metadata that the PREMIS metadata dictionary explicitly targets. Therefore, it seems likely that both standards would be applicable and useful for organizations, such as NSIDC, charged with both the preservation of and access to geospatial data. Moreover, these are precisely the recommendations of a recent study prepared for the National Geospatial Digital Archive (Hoebelheinrich and Banning, 2008).

7.3.3 PREMIS

The PREMIS metadata standard is designed to provide a core set of preservation metadata that is implementable across a broad range of contexts. The standard is accompanied by usage guidelines, recommendations, and tutorials for metadata creation, management, and use.

At the core of the PREMIS model is the concept of an intellectual entity – a "coherent set of content that is reasonably described as a unit, for example, a particular book, map, photograph, or database (PREMIS Working Group, 2005)." In this study, the intellectual entities correspond to data sets, where a data set is all of the files associated with a particular experiment or measurement, for example, all of the "MODIS/Terra Snow Cover 5-Min L2 Swath 500 m" products (Hall et al. 2006) or all of the "CLPX-Ground: ISA Snow Pit Measurements" (Cline et al. 2002).

PREMIS defines four other entities, each with defined mandatory and optional fields (see Fig. 7.2), required to preserve an Intellectual entity:

- *Object Entity* – the workhorse of the PREMIS model, storing the bulk of the preservation metadata about a particular object. Three types of objects are defined – bitstreams, files, and representations. A bitstream is a contiguous sequence of bits within a file that require separate preservation information, while a representation is typically a group of files along with associated structural metadata that together provide a description of an intellectual entity, in this case a data set, to be preserved. Objects may refer to other objects, events, or rights entities.
- *Event Entity* – records information about things that happened in the life of an object or a group of objects that could potentially affect its authenticity or accuracy.

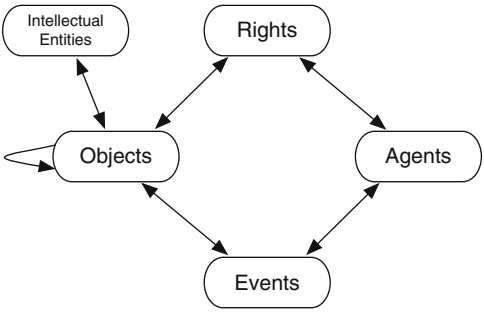

Fig. 7.2 The PREMIS data model, adapted from 2008, PREMIS data dictionary, Version 2. *Boxes* represent entities within the data model, while the *arrows* indicate relationships between entities that are supported by references within the metadata

- *Agents Entity* – records information about a person, organization, or software program that was involved with an event.
- *Rights Entity* – at the time of this study, the rights entity recorded information about the rights or permissions granted to an object or an agent. The rights entity was considerably redefined and expanded in Version 2 of the standard. It now records information about the intellectual property rights associated with an object and the associated permissions accorded the archive responsible for the object.

7.3.4 NSIDC Overview

The National Oceanographic and Atmospheric Administration (NOAA) chartered NSIDC in 1982 to house the national collections of the World Data Center A for Glaciology (http://nsidc.org/). In the years since its inception, NSIDC has taken on a variety of data management tasks with funding from NASA, NOAA, the National Science Foundation (NSF), and others. Currently the bulk of the data NSIDC manages are digital, ranging from small text-based "in-situ" data sets to large remote sensing satellite data sets. Two main systems are currently in use to manage and disseminate these data. The largest volume of data NSIDC manages is housed in NASA's Earth Observing System (EOS) Core System (ECS), which will not be described further here. The remainder of the data, some 15 TB or so and the majority of the data sets, are managed by custom systems developed by NSIDC.

A few years ago NSIDC developed a central catalog of all of its data holdings. Based on the NASA Directory Interchange Format (DIF) (NASA,, Directory Interchange Format (DIF) Writer's Guide, 2008), this catalog primarily contains information that the OAIS Reference Model would call descriptive and representational information. The catalog is used to dynamically generate Web pages for each data set, to support user searches for data sets, as well as to export summary metadata about holdings to other organizations such as NASA's Global Change Master Directory (GCMD) (http://gcmd.nasa.gov/). The NSIDC catalog supports multiple

data access methods that can be tailored individually for each data product. The catalog even supports the concept of "brokered data," which are data that are archived elsewhere, though NSIDC is funded to develop and maintain the metadata. However, the catalog does not maintain Packaging Information or Preservation Description Information. A variety of separate systems are used for data access and preservation. In many cases, these systems are manually intensive and because they are not integrated with the main catalog, issues of duplication of effort and potential for data inconsistencies exist. The Operations and Preservation Metadata Prototype at NSIDC is an attempt to rectify that situation by adding Packaging Information and Preservation Description Information to the NSIDC central catalog.

7.4 NSIDC Project

The NSIDC Operations and Preservation Metadata Prototype project had several goals:

- To prototype a uniform mechanism for recording packaging and preservation metadata for all of the data sets that NSIDC archives.
- To lay the groundwork for systems automation.
- To improve NSIDC data management efficiency and processes, hopefully a direct result of the successful completion of the first two goals.

The project started by mapping current high-level metadata to the components of the OAIS Reference Model. It was quickly apparent that the current catalog, along with NSIDC's standard set of documentation for a data set were generally equivalent to the OAIS Representational and Descriptive Information Metadata, but the preservation description and packaging information component of the model were lacking and required attention. Defining packaging information was delegated to the NSIDC operations group since they maintain the archive and have the greatest direct use for this information. A team consisting of representatives from each of NSIDC's major sponsored programs and the manager of NSIDC's archive concentrated on the preservation description. This latter team focused on the provenance, fixity, and reference components of the preservation description, since the contextual component historically had been defined through data set documentation. The definition of each component of the reference model was discussed in detail in the context of each of NSIDC's programs and candidate metadata items were drafted (see Tables 7.1–7.3). The results of these efforts led to a draft metadata model as well as a preliminary database schema (see Fig. 7.3).

Subsequently, NSIDC tested the draft model using several data sets from the CLPx project. In the midst of this exercise, the final report of the PREMIS working group was released. Since the PREMIS efforts and the NSIDC efforts closely paralleled each other, we decided to compare the PREMIS data dictionary to the NSIDC schema. While there are numerous minor differences between the two, overall the

Table 7.2 Candidate provenance information metadata items

Name	Definition	PREMIS field (if any)
External release ID	Version of the data released to the public.	Part of the object identifier
Archive version ID	Indicates a change to the data set significant enough to affect archival; but not large enough to warrant a public release (e.g., media migration or adding a new format browse file to the archive).	Part of the object identifier
Original date received	The earliest date at which data arrived at NSIDC for archival	Date time component of the beginning of an ingest Event
Migration history	Information about the migration history of the data set and where it is currently stored. For many data sets dating prior to the mid-1990s this information was held in a file folder, along with any agreements and correspondence with the donor	Content location plus a series of events – one for each migration event
Data modification history	Information about any modification made to the data (e.g., quality control checks flagging data as invalid, file reformatting, etc.). Again, much of this information was held in a file folder for older data sets	A series of events
Original provider name	The name of the person who originally provided the data to NSIDC	Agent name
Original provider address	The address of the original provider at the time they provided NSIDC the data – usually the address of the organization they worked at.	NA
Original provider organization	The name of the organization that the provider was associated with when the data was originally provided to NSIDC.	Agent name
Table of releases	The date of each release along with a description of the release	A series of events
Data agreement	Pointers to the document(s) that defines NSIDC's rights and responsibilities with regards to these data; the Level of Service to be supported; a summary of distribution rights and restrictions; and/or statements of ownership/custody rights	Permissions statement

Table 7.3 Candidate fixity metadata items

Name	Definition	PREMIS field (if any)
Fixity type	The type of mechanism used to guarantee that no undocumented changes to the files in this data set have occurred (e.g., MD5 message digest, Unix checksum)	Message Digest algorithm
Fixity description	A description of the exact mechanism used in detail.	

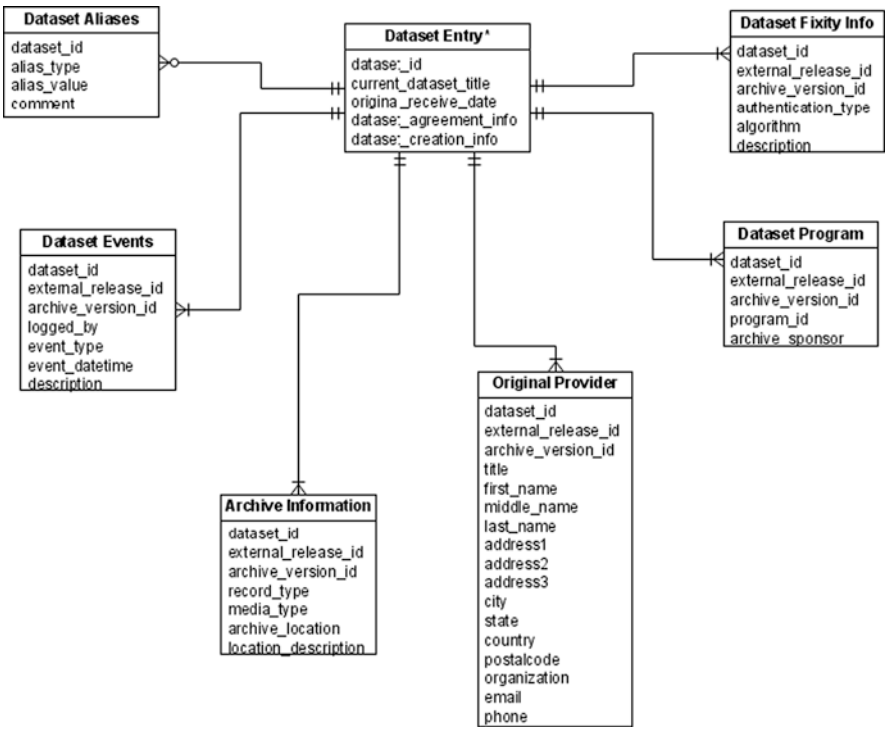

Fig. 7.3 Draft additions to NSIDC's existing catalog schema (* with the exception of dataset_id, only fields added due to this work are depicted)

agreement is very good. With the exception of a single attribute, it is possible to provide a mapping between all of the mandatory elements of the PREMIS data dictionary to items in the NSIDC schema and vice versa, albeit this is only done for the data set as a whole.

The PREMIS preservation level attribute and NSIDC's level of service are roughly equivalent, but require additional discussion before they can be directly mapped. We have not yet resolved whether there should be a unified or single level of preservation, or multiple levels associated with data sets managed by NSIDC.

Preservation of science data is more of a yes/no or binary switch rather than multiple level choices that objects in other organizations might employ. For example, with a complex document such as a Web page with an embedded movie, it may be sufficient to preserve the look and feel of the Web page without the ability to play the movie from within the page; however, the movie can be played outside the context of the Web page. Has the document as a whole been preserved? Well yes, but not completely. In the case of science data, it is generally considered essential that all the bits and bytes, ancillary documentation, and metadata be preserved in a useable form. It is the level of service provided to the user community that varies. For example, additional tools may be developed or maintained in order to facilitate data availability or alternately the level of support provided by NSIDC's User Services organization and scientists may vary.

It should be noted that many items in NSIDC's candidate metadata list include fields that are not applicable at the representation level according to Version 1 of the PREMIS metadata standard. The PREMIS standard treats these fields as only being applicable to individual files or bit streams within a file. At NSIDC, often all of the files associated with a data set share characteristics such as storage location, type of algorithm used to verify fixity, file format, etc. in common. Therefore it is useful to record these values only once for the data set and not to repeat these values for each file.

In the meantime, implementation of a prototype system for populating NSIDC's newly defined schema proceeded and was complete in May 2006. While the prototype never became operational, there are several points that can be made primarily from the comparison of the PREMIS and NSIDC schemata.

First, over the years an information hierarchy has arisen for science data. Data tend to be organized by sets where all of the data in the set has some type of commonality. For example, all the data in a set may have come from the same instrument, have been processed in the same way, have been acquired during a particular field campaign, or consist of measurements of the same quantity even though measured in a variety of ways. Unless the data set is small, the actual data in the set is typically organized into a series of granules, each of which consists of one or more files. The granules are the smallest unit for which metadata is managed. Data centers such as NSIDC, when developing data management systems, tend naturally to develop systems that preserve this hierarchy. For example, the catalog extensions that NSIDC worked on are at the data set level not at the file level. Below the data set level, granules and individual files are managed as part of the data set to which they belong. As a result, metadata about data sets and metadata about granules/files are separate in NSIDC's data systems.

In contrast, the PREMIS data dictionary generically accommodates all levels of information about items that need be preserved, and in NISDC's case, a data set and all granules/files in that data set. If the PREMIS data dictionary were directly implemented this would result in all dataset level information being stored in the same table as all granule and file level information. This could easily lead to scalability problems. While some data sets may only consist of one or a few files, data sets for long-lived satellite remote sensing programs typically contain millions of files. Tables containing tens and hundreds of millions of records could easily be expected to result even for small archives such as NSIDC. To avoid these scalability issues, NSIDC has chosen to preserve its hierarchical data structures.

Second, the PREMIS data dictionary has no concept of a version – metadata is explicitly assumed to be relevant to one and only one object and once that object is preserved, it cannot be changed. This may make sense from a library or cultural heritage archive perspective. New editions of a book replace older versions; copies of a painting are at best just copies if not outright forgeries. It makes less sense for digital data in particular digital scientific data. For science data sets, there is really only one significant property that needs to be preserved and that is the scientific integrity of the data. The integrity of the data is unaffected by many changes, up to and including changes in format. As long as the transformation is properly performed, a standard long held to mean that the transformation is reversible, the data

set has not been changed despite the factor that the digital files themselves have been altered. Reprocessing the data comprising a data set does not create a new data set; it merely creates a new version of the old one. More over, it is important to preserve information about the changes between the two versions of the data set. For example, scientists that used previous data sets may need to determine whether the results of the work based on the old version are still valid.

Third, NSIDC did not choose to implement the preservation rights section of the PREMIS data dictionary. With scientific data, if you have the data at all, you generally have the right if not the technical capacity to make as many copies of the data as you need to ensure data preservation. Of more importance for scientific data are distribution rights and preservation responsibilities. While distribution rights are outside the scope of the PREMIS charter, preservation responsibilities are not. Unlike the typical library, the holdings of a data center are usually unique and often very voluminous. So voluminous, that discussions are often held about the resource tradeoffs between storing a single backup copy of the data, or only storing a precursor data set and the processing software used to create the product and recreating the product on demand. Moreover, the organization responsible for the archive often changes over time, and as with any archive, removing a product from the archive is often necessary. While, responsibilities of this sort could have been defined for the first version of the PREMIS standard, the PREMIS group did not do so beyond noting that a Rules entity could have been defined. It also should be noted, that the second version of the PREMIS standard includes a much more complete handling of rights issues.

7.5 Conclusions

Our prototype PREMIS ops and preservation metadata tool has not been integrated into our operational environment. However, the development exercise has brought several issues into focus. Foremost, is recognition of the actual need to assemble and store preservation metadata as part of the data management process. We suspect the steps following our testing will include closer integration of the operations and preservation metadata with our discovery, collection, and inventory metadata systems. Discussion and ultimately resolution of levels of service definitions and procedures for application will merge with or clarify the utility of the PREMIS defined preservation level. Lastly, we confirm that, at least at the representation level, the PREMIS metadata standard appears to be useful and appropriate for digital earth science data.

Acknowledgement We wish to acknowledge the support of the NASA Distributed Active Archive Center Contract (NASA Contract NAS5-03099), the Arctic System Science Data Coordination Center at the National Snow and Ice Data Center (NSF/OPP/ARCSS Program Award ARC-0450901), and NOAA's data management efforts at NSIDC through the Cooperative Institute for Environmental Sciences (CIRES) Cooperative Agreement with NOAA (CIRES Cooperative Agreement, Task 4)

References

CCSDS 650.0-B-1: *Reference Model for an Open Archival Information System (OAIS)*. Blue Book (Standard). Issue 1. January 2002. [PDF (1,183,662 bytes)]

Cline, D., Armstrong, R., Davis, R., Elder, K. and Liston, G. 2002, Updated July 2004. CLPX-Ground: ISA snow pit measurements. Edited by M. Parsons and M. J. Brodzik. Boulder, CO: National Snow and Ice Data Center. Digital media. http://www.nsidc.org/data/nsidc-0176.html.

Cline, D., Elder, K., Davis, B., Hardy, J., Liston, G.E., Imel, D., Yueh, S.H., et al. 2003. Overview of the NASA cold land processes field experiment (CLPX-2002). In: Microwave Remote Sensing of the Atmosphere and Environment III, 4894:361–372. Hangzhou, China: SPIE, April 30. http://link.aip.org/link/?PSI/4894/361/1.

Federal Geographic Data Committee. FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee. Washington, D.C.http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf

Hall, D. K., Riggs, G. A., and Salomonson, V. V. 2006, updated daily. MODIS/Terra snow cover 5-min L2 swath 500 m V005. Boulder, Colorado USA: National Snow and Ice Data Center. Digital media. http://nsidc.org/data/mod10_l2v5.html.

Hoebelheinrich, N., and Banning, J. March 2008, An Investigation into Metadata for Long-Lived Geospatial Data Formats, Library Trends (in press), http://www.digitalpreservation.gov/news/events/ndiipp_meetings/ndiipp08/docs/session7_hoebelheinrich_paper.doc.

Hunolt, G. March 1999, Global Change Science Requirements for Long-Term Archiving. Report of the Workshop, Oct 28–30, 1998, USGCRP Program Office.

ISO, 2003, International Standard for Geographic Information – Metadata, ISO 19115:2003(E)

LOC, Schemas for PREMIS. PREMIS: Preservation Metadata Maintenance Activity, Library of Congress, http://www.loc.gov/standards/premis/schemas.html

NASA, Directory Interchange Format (DIF) Writer's Guide, 2008, Global Change Master Directory. National Aeronautics and Space Administration. http://gcmd.nasa.gov/User/difguide/.

NASA, Earth Science Data and Services Directory: Global Change Master Directory Web Site, Global Change Master Directory. National Aeronautics and Space Administration, http://gcmd.nasa.gov/.

NSF, August 2003. It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation, Final Report, Workshop on Research Challenges in Digital Archiving and Long-Term Preservation. Sponsored by the National ScienceFoundation, Digital Government Program and Digital Libraries Program, Directorate for Computing and Information Sciences and Engineering, and the Library of Congress, National Digital Information Infrastructure and Preservation Program, http://www.reference-global.com/doi/pdf/10.1515/MFIR.2004.15.

NSIDC, http://nsidc.org/.

PREMIS Working Group, 2005, Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group, OCLC and RLG, http://www.oclc.org/research/projects/pmwg/premis-final.pdf

PREMIS Working Group, 2008, Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group, Version 2, OCLC and RLG, http://www.loc.gov/premis/v2/premis-2-0.pdf

Chapter 8
Archive Standards: How Their Adoption Benefit Archive Systems

Robert H. Rank, Constantino Cremidis, and Kenneth R. McDonald

8.1 Introduction

Electronic archives have much in common: the information must be received, validated, catalogued, stored and made accessible to the consumer. This commonality increases even more as the types of electronic archives converge. In this section we will explore the benefits of employment standards for the development and operations of electronic archives. Although the discussion is applicable to a wide variety of electronic archives, we will focus our discussion in our experience developing and operating large archives of environmental data. We will discuss how the adoption of standards in general and a submission process developed using the recommendations for Space Data System Standards from the Consultative Committee for Space Data Systems (CCSDS) reference model for and Open Archival Information System (OAIS) has supported the development and operations of a large archive of environmental data.

8.1.1 Need to Archive Environmental Data

Organizations like the National Oceanic Atmospheric Administration (NOAA) have been collecting environmental data for over 30 years. Most of the times, these data are available to interested scientists and researchers via electronic archives. As the time span of the collection increases, the value of the data increases exponentially. This increase in value of existing data, coupled with the fact that over the next 15 years, current and planned remote sensing observing systems will produce volumes of environmental data on an unprecedented scale, makes very important for NOAA and other like organizations to develop electronic archives that can support the enormous increase in volumes. Additionally, if we add increased user expectations for data accessibility, archive inter-operability and easiness of data retrieval, we can see

R.H. Rank (✉)
NOAA/NESDIS/OSD/GSD/CLASS, Suitland, MD 20746, USA
e-mail: Robert.Rank@noaa.gov

L. Di, H.K. Ramapriyan (eds.), *Standard-Based Data and Information Systems for Earth Observation*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-540-88264-0_8,

how important it is to properly archive these environmental data in archives that are developed following generally accepted standards.

In NOAA's specific case, NOAA must make these data available to support a myriad of users. By the year 2017, plans for the current environmental observation satellite campaigns and numerous in situ observation programs will increase the total data volume (primary and back up copies) to more than 140 petabytes. This presents a particular challenge for long-term preservation and data discovery. Additionally, not all data will be held at a single archive (not even NOAA's data) therefore archive interoperability will be of significant importance.

The NOAA policy for acquiring, integrating, managing, disseminating, and archiving environmental and geospatial data and information obtained from worldwide sources to support NOAA's mission is established by NOAA Administrative Order (NAO 212–15) "Management of Environmental and Geospatial Data and Information".

Environmental data (as defined in (NAO 212–15)): Recorded observations and measurements of the physical, chemical, biological, geological, or geophysical properties or conditions of the oceans, atmosphere, space environment, sun, and solid earth, as well as correlative data and related documentation or metadata. Data may exist in either electronic or analog format.

Geospatial data (as defined in NOAA Administrative Order (NAO 212–15)): Information that identifies the geographic location and characteristics of natural or constructed features and boundaries on the Earth. This information may be derived from, among other things, remote sensing, mapping, and surveying technologies. Statistical data may be included in this definition at the discretion of the collecting agency.

8.1.2 Environmental Data at NOAA/NESDIS

The National Environmental Satellite, Data, and Information Service (NESDIS), a line office within the National Oceanic and Atmospheric Administration (NOAA), is responsible for archiving and disseminating environmental data collected by a variety of ground-based and space-based observing systems. The Comprehensive Large Array-data Stewardship System (CLASS) is NOAA's planned mechanism for securely archiving large-volume data and data products, and for making these data available to researchers, commercial users, and the public.

The ability to ensure on-going scientific stewardship for NOAA's environmental data and information will only be possible through extensive enhancement of NOAA's current data management for archives, which include data ingest, quality assurance, storage, retrieval, access, and migration capabilities. This goal will be met through the development and implementation of a standardized archive management system, which will be integrated with a robust, large-volume, rapid-access storage and retrieval system that is capable of storing the incoming large array environmental data, in situ data, and operational products as well as receiving a user's on-line data request, automatically processing the request, and providing the

requested data on the most appropriate media. CLASS will provide standardization in media, interfaces, formats, and processes for the very large datasets produced by satellites and radars. Additionally, CLASS will facilitate ongoing migration, preservation, and validation to new technology and media. CLASS is modular in design, built to integrate with automated real-time or near-real-time systems that deliver data. Transaction processing, including free online delivery of data, will be implemented to enable an essentially "hands-off" operation. Users that request data in physical media (tapes, CDs, DVDs, etc) pay to cover the costs of media itself, the service and shipping costs.

8.1.3 CLASS and NOAA's Environmental Data

CLASS is a critical capability and a key component of the infrastructure supporting NOAA's integrated observation and data system, providing permanent, secure storage, and safe, efficient Web-based data discovery to large-array data sets. It has been identified by NOAA/NESDIS as its primary Information Technology (IT) system to provide the archive, access and distribution capabilities for multiple NOAA data collections. NOAA will also examine the potential downstream expansion of CLASS infrastructure to include observing system data beyond its current intended use.

Currently, CLASS supports data from NOAA's Polar Operational Environmental Satellite (POES) and Geostationary Operational Environmental Satellite (GOES) missions. As stated earlier, new satellite and in-situ observation campaigns are being prepared for launch and operations. The volumes of data to be collected by these campaigns dwarf the data streams managed by the existing archive and distribution systems within NESDIS. The size, number, and frequency of data sets to be stored and distributed will require significant expansion of capacity for moving, storing, processing, and distributing data. This need for significant increases in capacity and performance brings two different but related types of challenges: one architectural, the other technical. As we will explain later, the solution to both of them is rooted in the adoption of standards.

To address the architectural challenge, CLASS has adopted the approach of developing an open archive that facilitates access and retrieval of data by a large and varied designated community. The collections will be held in multiple electronic archives and as we have seen before, interoperability among electronic archives is needed to facilitate user access to the environmental data that these archives house. To address the technological challenge, the need to process increasingly large volumes of data and the expected huge increase in the number of requests for these data. NOAA has directed CLASS to adopt the Open Archival Information System (OAIS) Reference Model to provide a framework and general guidelines for building the system, conducting its interactions with data providers and NOAA designated communities.

The benefit of using the OAIS-RM is that it provides a common set of functions, processes and agreements that are required to accomplish the data transfers and a

common terminology to establish the scope of the effort and the respective responsibilities of the data providers and the archive. The model identifies the need for a Data "Submission Agreement" (SA) between the producer and the archive as well as the need to clearly define the designated community for each of the archive data holdings. As we will present later, CLASS in particular and NOAA archives in general have clearly benefited from adopting the OAIS-RM. The model represents a synthesis of best practices across a wide range of archive programs and as a result is fairly high-level in its descriptions. CLASS had to make limited adjustments and tailor the OAIS-RM to meet its specific needs. However, adoption of the reference model has definitely assisted the project and accelerated its progress. The overall CLASS effort and its experience adopting the OAIS-RM as a standard provides an excellent case history in the use of a standards-based process in the design and implementation of a long-term archive.

8.1.4 CLASS Overview

Over the past few years, NOAA has decided to move towards an enterprise IT solution in support of NOAA's Archives. As part of this evolution, NOAA has identified CLASS as the system that provides IT capabilities to support NOAA's archive mission. CLASS is a NOAA/NESDIS initiative to develop and implement a single Information Technology (IT) system for the ingest, storage, access, and distribution capabilities for the NOAA National Data Centers (NNDCs). CLASS is currently hosted at multiple sites while providing a single interface to the consumer. This multiple-site capability is intended to improve system availability, scalability and enhance data integrity through replication at geographically disparate sites. CLASS is available at www.class.noaa.gov.

Under this enterprise vision, NOAA's archive responsibilities can be categorized into two distinct but closely related entities:

- Information preservers. The Information preservers are the NNDCs, the NOAA Centers of Data, and other NOAA offices with data preservation responsibilities.
- Supporting Systems. The supporting systems provide IT functional entities of an OAIS. These functional entities are primarily Ingest, Data Management, Archival Storage, and Access, as well as some aspects of Administration and Preservation Planning. CLASS is the principal but not the only supporting system.

8.1.5 ISO/CCSDS Standards for OAIS

In 1995 the International Organization for Standardization (ISO) asked the Consultative Committee for Space Data Systems (CCSDS) to develop a formal standard for the long-term storage of digital data generated from space missions. In preparation for this it became clear that a unifying framework to act as the

foundation for standards building activities was required. It was recognized that any reference model developed would solve cross-domain problems regarding the long-term preservation of digital materials. Consequently, the process of developing the model was opened to any interested individual or organization.

The OAIS-RM was developed through an open, iterative process of drafting, review, revision, workshops and community feedback. Draft versions of the reference model were released in May 1997 and May 1999. After review the draft ISO standard was published in June 2000. A final period of review and revision followed with final publication as ISO 14721 in 2003.

As indicated previously, in developing CLASS, NOAA is following the guidelines of ISO 14721.2003 that specifies a reference model for an open archival information system (OAIS). The purpose of this ISO 14721:2003 is to establish a system for archiving information, both digitalized and physical, with an organizational scheme composed of people who accept the responsibility to preserve information and make it available to a designated community. This Reference Model concentrates on three primary aspects:

- It defines the functions that an Archive organization must perform for supporting the requirements of an electronic archive, and long-term preservation
- It identifies the need to formalize the interactions between the archive and the producer. The OAIS-RM identifies a Submission Agreement as the mechanism by which a Producer and Archive reach agreement on how submissions will be accepted by the archive and what to do with them.
- It identifies the need to clearly define a Designated Community for the holdings of the archive. While this community may not include all potential users of the archive, it is the Designated Community who defines the access and dissemination requirements for the archive holdings.

A fourth, sometimes ignored, but maybe the most important of all is the development of a common terminology for describing functions performed by the archive. Before the development of the OAIS-RM there were many different terms for each element or function that comprises an archive and this led to many problems and misunderstandings.

8.1.6 OAIS: A Common Language for the Information Professionals

Although designed by space data curators, the OAIS model aims to be as context-neutral as possible. OAIS deliberately avoids jargon from both the IT and archival professions; this is very useful as it makes both groups speak the same language. Once acquired, the terms and language of the OAIS model enable the digital curator to communicate effectively with other national and international projects. The downside is that the jargon can deter those not yet immersed in “OAIS speak” and

act as a barrier to understanding and cooperation. The complexity of the OAIS reference model has led some practitioners to call for an "OAIS lite" which would make the model more accessible to smaller and less well funded institutions. We think that the activities of an electronic archive are so important for long term preservation of data, that any organization planning to build and operate an archive should spend the time to understand the OAIS. The OAIS can be divided into Functional and Information models.

The OAIS Functional Model ensures that:

- Data objects are appropriately Ingested, Archived and Managed.
- Administrative procedures are in place for the overall operation of the archive
- Planning for preservation takes place; including storage media and format migration planning, software decisions, implementing standards and creating ingest methodologies
- Data objects continue to be accessible and usable to those who need to use them.

The OAIS Information Model:

- Ensures that the necessary supporting information, (Metadata) to enable effective control and preservation of a data object, and record relationships between them, is collected or created.
- Ensures that any information needed to interpret a data object (Representation Information) is collected and assigned appropriately.

Of these two models presented in the OAIS, the most complex but also the most important is the Information Model. In the following paragraphs we will present a brief overview of the OAIS Information Model. However, we strongly recommend that any person that plans to build or is currently operating an Archive to read and understand the recommendations and guidance presented in the OAIS.

8.1.7 The OAIS Information Model: An Overview

In addition to defining the parties involved in the long-term preservation of digital materials, OAIS provides an information model for managing the digital materials as they pass through the system. A significant component of this model is the Information Package (IP). Each IP consists of:

- The digital object(s) to be preserved.
- The metadata required at that point in the system.
- The Packaging Information.

OAIS outlines three types of Information Package: Submission Information Packages (SIPs), the Archival Information Packages (AIPs) and the Dissemination Information Packages (DIPs). A brief description of these types of information

packages and identification of where adoption of standards would benefit the archive follows,

The Submission Information Package (SIP): At the SIP stage, the metadata accompanying the digital object is, ideally, supplied by the Producer who is generally the original creator of the material; in the case of personal archives it is perhaps more likely that a digital archivist working with the creator will provide the metadata. At this stage, the metadata will probably lack structure and may not be comprehensive at all levels of the archive. SIPs may also be supplied to an OAIS from another digital repository. Where another digital repository has supplied SIPs, the use of *interoperable metadata standards* will minimize the effort required to ingest the material into the new repository.

The Archival Information Package (AIP): At the AIP stage, the SIPs are prepared for preservation. During this process, the digital materials submitted for preservation are known as Content Data Objects and they are combined with the Preservation Description Information needed to administer their preservation. At this point the use of *data format standards and package formats standards* will minimize future data preservation efforts. OAIS breaks the PDI down into four sections:

- Reference Information: a unique and persistent identifier.
- Provenance Information: the history of the archived object.
- Context Information: relationship to other objects, e.g. the hierarchical structure of a digital archive.
- Fixity Information: a demonstration of authenticity, such as a hash value.

Additionally, OAIS also requires the archive to maintain the Representation Information required to render the object intelligible to its designated community. This might include information regarding the hardware and software environment needed to view the content data object. This is also known as Preservation Description Information.

The Dissemination Information Package (DIP): The DIP stage happens when a Consumer requests a digital object or group of objects from the OAIS. The OAIS supplies the object(s) packaged as a DIP comprising the object and relevant metadata. It is likely that the metadata accompanying the object at this stage will be more descriptive than technical; the Consumer is unlikely to want to see complicated metadata relating to fixity or representation. At this point the use of *metadata standards and data format standards* will make the DIP easier to understand and use by the consumer. In all instances, the Knowledge Base of the OAIS' Designated Community will guide the type and extent of metadata supplied.

8.1.8 Metadata Standards

One of the areas where standards are very important for the success of long-term data preservation activities of a digital archive is in the adoption of metadata standards. Metadata: data about data; is a fundamental element for future access and

utilization of the data. Just defining what metadata to collect is a fundamental and critical issue to resolve before the archive starts receiving any SIP. A very important decision for any Archive is to define a metadata standard and to carefully and fully adhere to it. The OAIS-RM recommendation points out that much of that information is "more easily available or only available at the time when the original information [science data] is produced" thus stressing the need for Producers to play an active role in creating and maintaining standards compliant metadata. Metadata are necessary for many purposes, such as calibration and data processing. This requires open standards that allow for a great degree of flexibility in data storage and retrieval. The experience NOAA/NESDIS is certainly consistent with this need. NOAA now finds it difficult to create metadata for older data holdings that were not well documented when they were originally provided to the data center.

CLASS is required to provide the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM), FGDC-STD-001-1998 and the Content Standard for Digital Geospatial Metadata: Extensions for Remote Sensing Metadata. FGDC-STD-012-2002 as well as Geographic Information Metadata (ISO 19115) compliant metadata that describe CLASS holdings. The CLASS submission agreements are developed is such manner that they clarify producer's metadata requirements. Those requirements must be addressed in standards compliant ways. For example, when contact information is provided, the FGDC contact template is used as an indicator of the required information. When parameters in the data products are described, those descriptions include the FGDC required information at a minimum. This approach certainly leads to improvements in the quality and consistency of metadata available to CLASS consumers and in the scientific usability of the data that CLASS provides. The following paragraphs provide a brief description of these standards:

Content Standard for Digital Geospatial Metadata: The standard was developed from the perspective of defining the information required by a prospective user to:

- Determine the availability of a set of geospatial data
- Determine the fitness of the set of geospatial data for an intended use
- Determine the means of accessing the set of geospatial data

As such, the standard establishes the names and definitions of data elements and compound elements to be used for these purposes, and information about the values that are to be provided for the data elements. The standard does not specify the means by which this information is organized in a computer system or in a data transfer, nor the means by which this information is transmitted, communicated, or presented to the user.

Content Standard for Digital Geospatial Metadata – Extensions for Remote Sensing Metadata: The purpose of these Extensions for Remote Sensing Metadata (hereafter Remote Sensing Extensions) is to provide a common terminology and set of definitions for documenting geospatial data obtained by remote sensing, within the framework of the FGDC (1998) Content Standard for Digital Geospatial Metadata (hereafter FGDC Metadata Content Standard or simply base standard).

Creating these Remote Sensing Extensions provides a means to use standard FGDC content to describe geospatial data derived from remote sensing measurements.

As described previously, the FGDC Metadata Content Standard was developed to define the information about a geospatial dataset required by prospective users. These Remote Sensing Extensions are to provide additional information particularly relevant to remote sensing:

- The geometry of the measurement process
- The properties of the measuring instrument
- The processing of raw readings into geospatial information
- The distinction between metadata applicable to an entire collection of data and those applicable only to component parts.

ISO 19115:2003: The purpose of this standard is to define the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data. The ISO 19115:2003 is applicable to: the cataloguing of datasets, clearinghouse activities, and the full description of datasets, geographic datasets, dataset series, and individual geographic features and feature properties. Though ISO 19115:2003 is applicable to digital data, its principles can be extended to many other forms of geographic data such as maps, charts, and textual documents as well as non-geographic data. ISO 19115:2003 defines:

- Mandatory and conditional metadata sections, metadata entities, and metadata elements.
- The minimum set of metadata required to serve the full range of metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data).
- Optional metadata elements – to allow for a more extensive standard description of geographic data, if required.
- A method for extending metadata to fit specialized needs.

8.1.9 File Format Standards

File formats are a crucial layer, indeed a hinge between the bits in storage and their meaningful interpretation. The proper access to and display of content depends entirely on the ability to decipher the respective bit stream, and therefore on precise knowledge about how the information contained within is represented. Consequently, file formats are one of the core issues of any digital preservation approach, and file format obsolescence is a major challenge for anybody wanting to preserve digital files.

Digital preservation has to guarantee the integrity, understandability, originality, authenticity, and accessibility of digital records and data. To enable this, preservation file formats have to fulfill a number of requirements. Their syntactic and

semantic specifications should be public, they should be free of patent and license fees, and ideally they are standardized by a recognized standardization body. Wide use and acceptance improve their long-term prospects. Preservation formats must be free of any crypto-graphical and compression techniques, their specification should be self-contained, and they should be storage media-independent. It becomes clear from the above that, generally speaking, open formats are to be preferred over proprietary ones, since they allow for unlimited use without license fees or patent issues, and the fully available documentation eases their future handling.

Since even the best, open and widely accepted file formats are not immune against technological obsolescence; migration to other, newer formats will be a necessity during the preservation process. Thereby, information will be transferred from one hardware and/or software configuration to another or from one generation of computer technology to a subsequent generation. Migration offers a number of opportunities. Since migrated documents are in an up-to-date file format, they are workable documents that can be accessed with current software tools, thereby lowering the user education need and the support costs. At the same time, the migration process can exploit advances in technology. Finally, continuous migration also entails refreshing of the support media.

However, there are some threats involved as well. Migration is not an established, uniform process, but always a highly specialized transformation of data, involving considerable human and financial input at an unpredictable rate. Also, every conversion carries the risk of data corruption, and subsequent migrations increase this risk. This holds especially true for cryptographic techniques such as digital signatures. Finally, the ever growing number of file format becomes all the more difficult to track. Despite of all these threats, migration is at present the only workable solution to preserve digital files for the long term.

8.2 Experiences Adopting Archive Standards

In this section we will move from the theoretic realm more into the practical realm. We will present our experience with CLASS adopting the archival standards presented in the previous sections. We will present what has worked, what did not and we will try to provide an explanation of why it did not. We will also present some of the most significant benefits that adopting archive standards has brought to CLASS. We have to add, that CLASS is very early in its development cycle, and we only expect the benefits to multiple and increase in variety as we move forward. Our experience with the adoption of archival standards has been nothing but positive.

8.2.1 A Little Background: Multiple NOAA Archives

Due to the diversity of the NOAA mission, NOAA information preservers exhibit significant variation with regard to mission, domains, management processes, and

designated communities. It is because of such diversity that it is not expected that there will be a single "NOAA Archive". Instead, there will be multiple "NOAA Archives", each performing its own archive administration and preservation planning activities while sharing the IT services of enterprise systems like CLASS. These various "Archives" will share some preservation planning and management policies and procedures. CLASS, as the NOAA enterprise IT system in support of NOAA's archives, is preparing to deal with this diversity of Archive management and preservation planning directives by establishing standards-based processes in general and by adopting the recommendations of the OAIS-RM in particular. One of the main elements from the OAIS-RM is the development of Submission Agreements. CLASS has developed numerous Submission Agreements in collaboration with various producers and NOAA Archives.

8.2.2 OAIS-RM Submission Agreements

A Submission Agreement is an agreement between a producer and an archive on how the data will be submitted to the archive for preservation and dissemination to the designated community. Submission Agreements include information on:

- Human contacts (technical, metadata);
- Designated community
- Data access and dissemination options
- Producer to Archive data transfer protocols
- Validation tests; errors conditions and actions
- Data formats and standards
- Metadata
- Data quality information
- Lineage and other data Parameters
- System performance

Various elements of the Submission Agreement are IT-related and under NOAA's conceptual archive architecture, CLASS is responsible for those IT components while Archive related responsibilities are conducted by different organizations.[1] Development of a Submission Agreement is a three-party effort: The Producer, the Archive and CLASS. This adds an additional complexity not envisioned by the developers of the OAIS-RM. It is under this reality, that CLASS is developing various new policies, processes, and procedures to address these complexities.

[1] In general this responsibility falls to one of the NOAA Data Centers: the National Climatic Data Center (NCDC) in Asheville, NC; the National Geophysical Data Center (NGDC) in Boulder, CO; or the National Oceanic Data Center (NODC) in Silver Spring, MD. The three data centers are collectively known as NOAA National Data Centers (NNDCs)

In particular, CLASS has developed and implemented a number of standard processes to aid in the development of Submission Agreements. The five most important are:

1) The development of a Submission Agreement template. This is a comprehensive and fully annotated template that provides clear examples for all sections of the Submission Agreement document. This template is frequently reviewed and updated to ensure usability, clarity, and understanding.
2) Establishment of a single Point of Contact (POC) for the producer with the Archive. One issue that causes frustration from the producer's perspective was having to deal with multiple people during the development of a Submission Agreement – and thus having to repeat concepts every time the submission agreement reaches a milestone causing a POC to change. CLASS has established a "cradle to grave" approach when developing a new Submission Agreement. A Submission Agreement is assigned to one person who is fully responsible for coordinating with the producer and CLASS for the completion and approval of the Submission Agreement.
3) Establishment of monthly Submission Agreement status review meetings among CLASS and the representatives of NOAA archives who are working on Submission Agreements. These coordination meetings ensure that CLASS and the Archives are aware of which Submission Agreements are under development, what the pertinent need dates are, and what the producers' and designated community's expectations are, among other issues.
4) Establishment of a Metadata Management Repository and a Metadata Manager. A single system and a single organization, the NOAA Metadata Management Repository (NMMR), is responsible for gathering "collection level metadata" following the agreed upon metadata standard. Ensuring is this way uniformity of format and content. This single organization interprets the metadata standards and provides guidance on which information, in what format, and to which level of detail must be collected in the Submission Agreement.
5) Improvements to the requirements management process. It was noted that during the development of Submission Agreements, especially when drafting the sections for data access and dissemination, additional functional requirements were sometimes being included. Support for these access and dissemination requests would mean that the IT system supporting NOAA archives would need to be enhanced, in some cases at significant cost and/or with little forewarning. CLASS recognized that the development of a Submission Agreement in some instances had the unintended consequence of circumventing the requirements vetting and approval processes already in place. Enhancements to the Submission Agreement review and approval process now include: clear and early identification of proposed new requirements; compilation of these requirements into an appropriate document; and submission of that document through the standard requirements review and approval process. These steps are necessary before proceeding with review and approval of the Submission Agreement that brought the need for these requirements to light.

By implementing these five process improvements for the development of Submission Agreements, CLASS in particular and NOAA's archives in general have found that many of the early adoption difficulties have been overcome and that the process of developing Submission Agreements is smoother and better accepted by producers. There is also a notion that it is generally wise to tell the Data Producer that the Archive is expecting to follow the OAIS Reference Model in its operations and that there is a formal negotiation of content, roles, and responsibilities that will follow the initial contact.

8.2.3 Need for Interface Control Document

Groups of NOAA's data producers, especially its satellite data processing systems, tend to share data distribution points. A diverse group will typically share a single data production and distribution system. This is true for the current Environmental Satellite Processing System (ESPC) where different producers share the same IT systems; and will be true in the future for the National Polar-orbiting Operational Environmental Satellite System (NPOESS) Data Exploitation System (NDE) and for the NPOESS Interface Data Processing Segment (IDPS), among others. Under these circumstances, it is impractical to include in every Submission Agreement the technical characteristics of the shared interface through which the data will be submitted to CLASS for archival storage. The reasons are various, the most important are:

- One change to the interface would require changes to multiple documents;
- Producers do not have authority to negotiate the interface requirements for this IT system
- Only CLASS has authority to negotiate CLASS' interface requirements.

To address these problems, CLASS implemented the development of an Interface Control Document (ICD) for documenting the technical characteristics of each interface. Although this document is not called upon by the OAIS-RM, it is important an extension to the model that has significantly helped CLASS meet its requirements for receiving data. CLASS and the representatives of the IT system that will submit the data to CLASS develop and control this ICD. The ICD describes:

- Data transfer protocols;
- Data transfer validation mechanisms;
- Error conditions and recovery mechanisms;
- Data transfer volumes and performance requirements;
- Network architecture; and
- System security considerations.

The ICD is developed by and for the IT systems. It is written without representation of members from the NNDCs (as part of the NOAA Archives) and

the producers. Although, in theory, one Submission Agreement can be related to multiple ICDs, our experience, so far, has been that the opposite is true – one ICD relates to multiple submission agreements. These two documents, a Submission Agreement and its corresponding ICD, are related by the data that they document.

8.2.4 Benefits of Adopting Archive Standards to the Producers

Although CLASS' experience adopting Archive standards is generally very positive, its benefits do not stop at the archive doors. The adoption of archive standards by the Archive, benefit also the producers and the consumers. In this section we will address the benefits to the producers.

During the development of Submission Agreements for new data that will be submitted to CLASS for storage, archival, and dissemination, CLASS' experience is that producers tend to be, understandably, more interested in resolving issues related to algorithm development, data production, and dissemination to real-time users than in working with NOAA Archives to develop a Submission Agreement. It is because of this that NOAA Archives have found it necessary to clearly state the goals and benefits to producers for development of Submission Agreements. Without full participation from the producer, CLASS has found that development of a Submission Agreement takes a prolonged amount of time (years) to complete. The following are the primary benefits that a producer will receive by developing a Submission Agreement:

Early identification of required metadata: During the development of Submission Agreements, all collection and granule level metadata that will be utilized by the archive to catalog, provide access to, and disseminate the data to the designated community must be identified. Required metadata is clearly identified, helping the producer minimize last minute changes or scope creep, by asking the following questions:

- How will the data be cataloged?
- How will the data be discovered and searched?
- How will the data be disseminated to the designated community?
- What additional information is valuable to the designated community?
- What additional information is required for reprocessing of the data?

Early identification of data that needs to be submitted to the archive: This benefit is primarily a consequence of the previous benefit. The full needs of which data should be archived are quickly identified by asking questions like:

- What is the primary use of the data by the designated community?
- What information is required for reprocessing the data?
- What companion data is needed for understanding the data?

It is CLASS' experience that it is very easy to fall into the trap of saying that "all" data will be submitted for archival without clearly defining what "all" means. Development of a Submission Agreement is the surest and most efficient way of clearly and completely defining "all data". The development of the Submission Agreement has the additional benefit of defining the interface performance requirements and therefore supports system capacity planning activities. CLASS has identified many other benefits to the producer. The two presented here are simply the ones with the most significant beneficial impact for the producer.

8.2.5 Benefits of Adopting Archive Standards to the Consumers

Standardization allows consumers to have confidence in the quality and reliability of the products and services. Archive standards – whether they are for data collection, data transfer, documentation (metadata), or for software – are all designed to facilitate the dissemination, communication, and use of information by multiple producers and users. Almost all standards either rely on or incorporate metadata in order to accomplish their purpose, making the adoption of metadata standards an extremely critical decision for an archive.

Consumers directly and significantly benefit by the archive adoption of standards. In this section we will present the most significant ones.

- Data Interoperability: One the FGDC's primary goals is to provide a consistent means to directly compare the content and positional accuracy of spatial data obtained by different methods for the same point and thereby facilitate interoperability of spatial data.
- Common Vocabulary: Standards provide a common set of terms. With standards, there is no confusion about what is being communicated by a particular term, from one metadata record to the next, the terminology is the same.
- Easier Access to Information: Standards allow for quick location of a certain element. If a standard is used, finding a specific piece of information in a metadata record will be much easier than if no standard is used.
- Automation: Standards enable automated searches; when standards are used, computers can be programmed to search and find useful data sets. This function of standards will become more important as more electronic data clearinghouses are built.
- Stability: Some standards are federally mandated. Under Executive Order No. 12906, all federal agencies and organizations receiving federal funds must document their geospatial data using the Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata. Conformance to this standard minimizes duplication of effort in the collection of expensive digital data; fosters cooperative digital data collection activities and establishes a national framework of quality data.

8.2.6 Use of Archive Standards – Summary of Benefits and Challenges

So far we have presented how a complex and diverse organization such as NOAA greatly benefits from the adoption of standard processes and procedures in general and from the adoption of the OAIS-RM in particular. CLASS, as the enterprise IT system in support of NOAA archives, has embraced the OAIS-RM and its recommendations. One such important recommendation is the development of Submission Agreements between the producer and the Archive. We also presented the unique challenges that NOAA and its Archives face when developing Submission Agreements. These challenges include the fact that more than two organizations develop and approve the Submission Agreements, and that producers may share a single IT system for submitting their data to the Archive.

The use of the OAIS Reference Model has served CLASS in a number of ways. In what might first appear to be a simplistic example, the reference model has given the data provider and the archive a common terminology with which to frame the discussion. It has also identified a common set of functions that are required to archive data and information and a set of processes to establish the specific requirements that are associated with those functions. Further, it identifies a set of documentation to capture and record those requirements and specifications. Although some of these functions and processes have had to be tailored and extended to meet the specific needs of the CLASS project, the reference model has provided an excellent foundation on which to build.

One source of difficulty in the direct application of the OAIS Reference Model to the long-term archive of NOAA Environmental data in CLASS is that in some cases (for historical data), the data provider is not the original data producer but rather is an existing, operational archive itself. This requires the additional step of mapping and adapting the conventions of Provider-Archive to the OAIS model being used by CLASS. Another challenge in the long-term archive of NOAA Environmental data is the shear magnitude, diversity, heterogeneity and complexity of the science products and the issues associated with the decisions on how to organize the information that accurately captures and conveys that complexity.

In some of the projects for preserving historical data, the data preservation problems that we will face in the future are already becoming evident. Deciding what data to give preference is a challenge by itself. Lately, the decision has been made to initially transfer the lower level instrument data; however, to use that data to generate higher level science products requires orbit and attitude data and significant details in each metadata record. Additionally, processing software and all of the information on the computer environment in which it runs must also be stored in the archive. Finally, for true data preservation all of the expertise that is required to generate and interpret the data and associated products must also be captured and preserved. CLASS is still working to solve some of the complexities related to historical data rescue projects and learning from them to improve its data preservation policies and practices. Electronic Data Preservation is a research area for which we hope will have significant developments in the near future.

Chapter 9
An Association Rule Discovery System Applied to Geographic Data

Laura C. Rodman, John Jackson, and Ross K. Meentemeyer

9.1 Introduction

Geographic information processing and spatial data analysis activities have increased dramatically in recent years, due to improvements in Geographic Information Systems (GIS) technologies and an explosion in available data sets. Many of the data layers used in geospatial analysis are derived from remote sensing data products. These data sets might include vegetation conditions, land use, human structures, and terrain. In some cases, the results from an association rule analysis can be used as ancillary information to assist in the interpretation of remote sensing images. Associations between data layers can be used to guide image recognition and identification (King 2002), or can be applied to validation and error checking of data. Association rules can also be applied to prediction, in which rules found in one domain are applied to new domains where the data are not complete. In those cases it may be useful to infer the presence of features in an area based on the known patterns of occurrence elsewhere.

Data mining techniques are promising for their ability to detect patterns in large data sets, and they have gained wide usage in the analysis of business transactional data. However, spatial data differ from transactional data. The spatial influence of features on one another, based on position, orientation, and proximity, must be accounted for. Spatial data may also contain both numeric and categorical data, and there are a variety of data formats used for both vector and raster data types. Geographic data sets can be very large, and data may exist at different scales, resolutions, geographic coordinate systems, and projections. Thus, data mining techniques appropriate only for numeric field data or for transactional data will not be sufficient for spatial data mining.

A great deal of research into spatial data mining has occurred in recent years, yet most of these advances have not found their way into an easy-to-use tool for mainstream GIS users. Many geospatial analysts, especially those with limited statistics

L.C. Rodman (✉)
Nielsen Engineering and Research, Inc., Mountain View, CA, USA
e-mail: rodman@nearinc.com

L. Di, H.K. Ramapriyan (eds.), *Standard-Based Data and Information Systems for Earth Observation*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-540-88264-0_9,

expertise, could benefit from a tool to find multivariable association relationships in a geographic data set. The goal of this work is to develop a software tool for discovering association rules in geographic data sets. This tool will integrate with established GIS software, work with standard geographic data formats, and have an intuitive graphical user interface. The resulting system will work with both numerical and categorical data types, and will discover rules composed of distinct geographic features, continuous numerical conditions, their locations and spatial relationships, and the strength of their associations. The specific application is the discovery of sets of spatial features or attributes that tend to occur together, with a certain probability.

Association rules describe the coexistence of objects or conditions within a data set. An association rule expresses the statement A => B, or in words, the existence of a set of objects A implies the existence of a set of objects B (Agrawal et al. 1996). With geospatial data, the object sets A and B can include distinct geographic features, an instance of a continuous environmental condition, and the spatial relationships between the features and conditions. A rule is not limited to associations at a single geographic location, but can also include objects that are within an influential distance of one another. Association rules are postulated from the variables in the data, and the strength of the association rule is measured by its *support* (the number of co-occurrences of A and B in the data set) and *confidence* (the frequency that an instance of A also contains an instance of B).

Methods for the discovery of association rules have been investigated over the past decade (Agrawal et al. 1996, Hipp et al. 2000), and a few studies have applied them to spatial data (Koperski and Han 1995). A general problem in association rule discovery is computational efficiency, since the number of combinations that can be tested increases exponentially with the number of variables in the problem. Another issue is how to limit the resulting rules to only those that are interesting or useful. Other difficulties arise in the application to spatial data, such as the need to handle geographic data formats and the need to relate both overlapping and non-overlapping features. These issues are all addressed in this work. Various methods are employed to limit the number of rule combinations that are examined in a multivariate problem, including a statistical method and an algorithmic technique. In addition, the software design and the user interface enable the user to apply expert knowledge to the selection of rule variables to keep them relevant. The data mining software integrates with a GIS to handle tasks such as data and coordinate conversion and geoprocessing. User feedback was solicited throughout the development process to ensure that the software is relevant to the GIS analyst community.

9.2 Association Rules in Geographic Data

The goal of this work was to produce a software system to perform association rule data mining in geographic data sets. The resulting software is called *Aspect* (Associations in Spatial Data). To accomplish this task, a methodology

was developed to read in geographic data sets, to structure the data according to level of detail, to perform various analyses to identify the significant variables in a relationship, and to compute the strength of candidate association rules.

This section describes the concepts used in the development of *Aspect*. The definitions and building blocks for spatial association rules are discussed first, followed by the data characteristics and rule formation. The final two subsections describe the sampling methodology used to acquire the individual data points and the various methods used to identify the significant variables to include in a rule.

9.2.1 Rule Characteristics

9.2.1.1 Spatial Association Rules

Association rules are rules in the form A => B ("A implies B"), where A and B are sets of objects or other variables in a problem. The right-hand side of the rule is called the *consequent*, and the left-hand side of the rule is called the *antecedent*. Association rules deal with the relationships between specific conditions or variable instances rather than with variable trends; thus, they are appropriate for data that are nominal or categorical in nature. A great many geographic data are categorical; for example, distinct features such as structures and water bodies, and descriptive variables such as vegetation types and soil types. Other variables are numerical, such as elevation, average temperature, and slope. In an association rule, numerical values are handled by binning them into categories. Thus, distinct geographic features, categorical variables, and (categorized) numerical values can all be considered as members of sets A and B. A rule is not limited to associations of overlapping variables, but can also describe objects that are spaced apart within an influential distance of one another. In that case, the spatial relationship between the objects is also included in the set A or B.

Since geographic association rules deal with distinct features, categorical data, or discrete groupings of numerical data, it is natural that the data mining software deal primarily with vector data types. These data types may have one or more attribute name/value pairs associated with them. The association rules that are examined can be general (containing geographic objects without regard to attribute values) or more specific (multiple rules, each containing the same geographic object but with different values of the attributes).

Association rules are postulated from the variables in the data set, and the strength of each rule is measured by its *support* **P(A, B)** (the probability that both A and B occur together in the data set) and *confidence* **P(B|A)** (the probability of B given A). The support can be written two ways, as the absolute number of occurrences of A and B together, and as the absolute number normalized by the total number of data points. The confidence is the support number normalized by the number of occurrences of A. Other measures of a rule's strength are occasionally used, such as the *lift* **P(B|A)/P(B)**, which is a measure of correlation (Han and Kamber 2001). The lift is the probability of B given A, normalized by the probability

of B regardless of A. If the lift is greater than 1.0, it means that the presence of A increases the probability of B over the probability of B occurring in general. Conversely, if the lift is less than 1.0, the presence of A decreases the probability of B, so the relationship A => B is considered to be incorrect. The higher the lift (over 1.0), the stronger the relationship A => B. Minimum thresholds for the support and confidence are specified by the user for a particular problem and may be modified as the analysis progresses; in particular, a tradeoff is often needed between the threshold values and the resulting number of useful rules (Scheffer 2001).

9.2.1.2 Horizontal and Vertical Rules

Two types of spatial association rules are considered in this work, *vertical* and *horizontal*. Vertical relationships are those that occur between features or conditions at overlapping geographic locations, and horizontal relations are those between features that are spaced a distance apart. Geographic entities are known to influence other nearby entities, and the influence decreases as the distance between the features increases.

Vertical rules are the more common situation, where all the variables of interest overlap in space. The following shows an example of a vertical rule (the symbol "^" means "and").

AvgMaxTemp(62-66)^AvgMeanTemp(53-55)^AvgMinTemp(43-44)
^Elev(500-1000)^Precip(38-44) => Redwood
Support = 28, Confidence = 77%, Lift = 6.48

In this rule, the climate conditions (average temperatures and the precipitation), elevation, and vegetation class (Redwood) are all sampled at the same locations to gather the statistics. The conditions described in the rule were found at 28 separate locations. The spatial relationship "Overlap" is implied rather than stated explicitly in the rule.

Conversely, horizontal rules must include the spatial relationship between the non-overlapping variables. The relative locations of the geographic features are also given explicitly. Examples of horizontal rules are:

Feature_at(X, Shopping_Mall) =>
Feature_at(Y, Downtown(economy, poor))^Distance(X, Y, 1 - 5)

Feature_at(X, Feedlot) => Feature_at(Y, Urban) ^ Distance(X,Y, > 5)

Feature_at(X, Ocean) =>
Feature_at(Y, Soil(type, sand)) ^Adjacent(X, Y)

Feature_at(X, River) =>
Value_at(Y, Slope(value, < 10%)) ^ Adjacent(X, Y)

The *predicate* (a semantic expression denoting a property or relationship) "Feature_at(X, object name)" means that the geographic feature "object name" exists at location X (the same holds for Y). The first example rule above states that shopping malls are associated with depressed downtown economies if the mall is located between one and five miles away from downtown. The second rule states that urban areas are typically found more than 5 miles from feedlots. The third rule states that the soil type adjacent to an ocean is sand. The fourth rule states that the slope on a river bank is less than 10%. These horizontal rules include the "Distance" and "Adjacent" spatial predicates that relate locations X and Y. Note that the spatial predicates "Distance" and "Adjacent" are always placed in the consequent (right-hand side) of an association rule.

More complex horizontal rules can also be formed, such as a rule that combines the "Overlap" and "Distance" relationships. The following example shows the relationship between bridges and road/water intersections depending on the distance from an urban area.

$$\text{Feature\_at(Y, Road - Water Intersection)} => \text{Feature\_at(Y, Bridge)} \wedge \text{Feature\_at(Z, Urban area)} \wedge \text{Overlap(X, Y)} \wedge \text{Distance}(X, Z, 0\text{-}20)$$

9.2.1.3 Data Format

The data used in the association rule discovery system is geographic data in a vector format. The data describes either point, line, or polygon shapes, and each shape has optional attributes associated with it. The overall feature type has an *object name*, such as "Road," "Vegetation Type," or "Precipitation." *Attribute names* describe any attributes that belong to the object name, such as "width" or "Primary vegetation." *Attribute values* refer to the possible values for a given attribute name, such as "4-lanes" or "Redwood." Thus, there is a hierarchy to the data format, where *object name* is the most general, *attribute name* is in the middle, and *attribute value* is the most specific.

9.2.1.4 Predicate Format

As seen in the rule examples above, the components of a rule are written in a predicate format. The geographic data in vector format are converted to predicates for the rule analysis. For vertical rules (where location and spatial information is not stated explicitly), the predicates describe the object names and any relevant attribute names and values. For horizontal rules, the most common predicate format describes the relationship between a geographic object or condition and its location. Discrete objects or categorical variables are referred to as "features." Numerical conditions (such as environmental variables) are referred to as "values." Thus, the two general types of geographic object predicates used in the association rules are:

Feature_at(X, obj_name(attrib_name, attrib_value))
Value_at(Y, obj_name(obj_name, obj_value))

X and Y refer to a geographic location. Thus, the first predicate states that a discrete object with a given name (obj_name) is found at location X. If the object has attribute values associated with it (this is optional), then the attribute name/value pairs are added in parentheses after the object name. The second predicate states that a continuous condition with a given name/value is found at the location Y. This data must also be in vector format, and typically the object name and attribute name provide redundant information, as there is only one value associated with the object. Examples of these predicates are:

Feature_at(X, Road(surface, paved))
Value_at(Y, Elevation(elev, 1000))

The first example states that a road exists at location X, and the surface of the road is paved. The second example states that an elevation of value 1000 exists at location Y. Rules can contain any number of geographic object predicates.

Other predicate formats may be used as a component of a rule. In addition to describing relationships, predicates can also describe properties of an object. For example, a predicate may be used to describe a feature density, length, or area. Examples are Density(Road,1000 units) and Area(Orchard, 25 units) (density and area units are specified by the user of the analysis software, for example, length of roadway per 30×30 m grid cell).

In addition to geographic object predicates, the program also uses spatial predicates. These predicates describe the spatial relationship between two or more geographic locations. The spatial predicates used currently are:

Overlap(X, Y)
Distance(X, Y, range)
Adjacent(X, Y)

Again, X and Y refer to locations. The first predicate states that X and Y are at the same location (they overlap). The second predicate states that the two locations X and Y are a specified distance apart. The distance is given as a range rather than as an exact value. The third predicate states that X and Y are adjacent to each other.

In the case where all the variables (features and values) in a rule overlap in space, the rule may be simplified by omitting the X and Y locations and the Overlap(X, Y) predicate. Thus, the following rule

Feature_at(X, Road(surface, paved)) =>
Value_at(Y, Slope(value, $< 15\%$)) ^ Overlap(X, Y)

may be simplified to

Road(surface, paved) => Slope(value, $< 15\%$)

There are two different ways to handle multiple attributes of a single geographic object predicate. The attributes can be all placed into a single predicate, or they can each be placed into a separate predicate. An example of the first choice is

Feature_at(X, Road(surface, paved,)(width, 2 lanes))

An example of the second choice is

Feature_at(X, Road(surface, paved))
Feature_at(X, Road(width, 2 lanes))

The choice about whether to combine attributes into one predicate or split them into multiple predicates is problem-dependent. Combining the attributes in the first example is the best choice if the user is interested in relating roads with non-road objects within an association rule. Splitting the attributes is the best choice if the user wishes to investigate relationships between road surface types and road widths within an association rule.

9.2.1.5 Variable Hierarchy

As mentioned above, the data format consisting of *object name*, *attribute name*, and *attribute value* forms a hierarchy from most general to most specific. The geographic data is entered, and variables from each of these levels are possible candidate variables for association rules. The candidate association rules (those that will be tested for strength) can consist of variables from any level. If a variable from the *object name* level is chosen, then the candidate rules will include all possible combinations of the attribute name/value pairs that are associated with that object name. Thus, if there are a lot of attributes, there could be a large number of rules examined. (If only the object name is of interest, with no attributes, then the attribute data can be omitted when the data is loaded from the geographic data files into the program). If a variable from the *attribute name* level is chosen as a component of a rule, the candidate rules will be limited to those containing the associated object name and only that attribute name. All attribute values for that attribute name will be examined; thus, there may still be multiple rules to evaluate. Selecting an *attribute value* as a component of a rule will limit the candidate rules to only that single attribute value for an object name.

This method of organizing the data according to a detail hierarchy was found to work best in practice. The data is also structured according to categorical/numerical data types, using the "Feature_at" and "Value_at" predicates. Also, the ability to group together or separate attributes into predicates provides problem-dependent structure to the data.

9.2.1.6 Binning Attribute Values

The association rule analysis and the contingency table analysis described below both require frequent samples of data to draw conclusions. A strong rule depends

upon a high support count, and the contingency table analysis is not valid unless certain frequency requirements are met. If an attribute value does not occur often enough in the data set to meet these requirements, then it can be combined with one or more other attribute values into a new category or "bin." Then the combined values can be considered in the association rule and the contingency table analysis. Both categorical attribute values and numerical attribute values can be binned.

As an example of a bin, consider a vegetation data set with an attribute name "Vegetation Type." The attribute values include "Valley Oak Woodland," "Coastal Oak Woodland," and "Blue Oak Woodland." If these three attribute values do not occur individually with sufficient frequency, then they can be combined into one "Oak Woodland" bin. If the combined bin has sufficient samples, then that bin can be used in the contingency table and in the association rule instead of the individual attribute values.

9.2.2 Rule Formation

All of the spatial association rules are composed of predicates formed from object names, attribute names, and attribute values. However, the rule formation process differs for vertical and horizontal rules. This section describes these two processes.

9.2.2.1 Vertical Rules

Often during an analysis, the user is interested in testing a number of different rules to see which ones are strongest and of most interest. For example, if a user wishes to investigate rules of the form: vegetation type => soil type, and there are five vegetation types and five soil types, then there will be 25 different rules to test to try to find a strong relationship. In other cases, there may be a large number of potential predictor (antecedent) variables in a data set, and the user may not be sure which subset to include in the rule. In that case, it may be of interest to test different combinations of antecedent variables to see which result in the most interesting rules. Vertical rule formation is designed for these types of situations.

For a vertical association rule, the consequent (right-hand side of the relationship A => B) is limited to one variable. Any number of variables can be in the antecedent. It is assumed that the user wishes to predict the consequent based on the values of one or more variables in the antecedent. Since the variable in the consequent is of the most interest, in *Aspect* it is called the "primary variable." Every association rule has one and only one primary variable. The variables in the antecedent are referred to as "secondary variables." Every association rule has one or more secondary variables.

To form an association rule, the user can select candidate primary and secondary variables from the hierarchical list of object names (more general), attribute names, and attribute values (more specific). If a more general variable is selected, a larger number of potential rules will be tested since all of the combinations of attribute names and values associated with the object name need to be examined. If a more specific variable is selected, then fewer potential rules will be tested.

The primary variable is a variable that the user would like the rules to contain as a minimum. For example, if the user is seeking a rule that associates predictor variables with soil type, but does not know in advance the predictor variables, then the primary variable will be "soil type" and the secondary variable(s) can be found by the software. This means that the rule output will contain soil type and any other combination of variables that creates a strong rule. If the user has no primary or secondary variables in mind, the program will select them automatically by taking all variables from a given hierarchical level (the level is chosen by the user).

If there are a large number of variables to consider, the user can perform a contingency table analysis (described below) to narrow down the choices. Once the final list is selected, *Aspect* will form candidate rules using all combinations of one primary variable and one or more secondary variables. The candidate rules are then tested for strength using the support and confidence measures, and any rule that meets the minimum thresholds for these measures is output as an association relationship.

9.2.2.2 Horizontal Rules

For horizontal rule analysis, it is assumed that the user knows in advance which variables to include in a rule, and the number of variables that can be used with a spatial relationship is limited. For horizontal rules (in particular, for the relationships "Adjacent" and "Distance"), variable conditions at only two non-overlapping locations can be considered. The two locations are related to each other by the spatial predicate. Any number of additional variables can be included in the rule, but these variables must overlap each other at one of the two locations referred to by the spatial relationship.

For horizontal rules there is no process of narrowing down a list of variables to the most significant as there is for the vertical rule analysis. Thus, horizontal rule formation does not use the concept of primary and secondary variables. Instead, the rules are built up from their individual components (both geographic object predicates and spatial relationship predicates). The spatial predicates are automatically assigned to the rule consequent, and the user can specify the geographic object predicates to also include in the consequent. The rule component build-up method is also used for cases where a variable is modified spatially, for example, by computing its density (occurrences per unit area).

9.2.3 Sampling

9.2.3.1 Vertical Rule Sampling

A sampling scheme is needed to collect the data points used to compute the support, confidence, and lift of a rule. Multiple sampling schemes were investigated during the course of this work, including several schemes to sample individual polygon features. However, these techniques were found to be problematic (Rodman and Jackson 2007). It was determined that a uniform grid sampling scheme is the best

approach to use for polygon layers, in terms of generality and the ability to automate it. The resolution of the grid can be controlled by the user to be fine enough to pick up the features of interest while reducing the possibility of introducing bias due to oversampling. For typical domains a 30×30 grid is sufficient. Within each grid cell a sampling point is chosen at random.

The following techniques are used for sampling overlapping data:

1. All data types are point data.
   An example of this would be the relationship between road/water intersections and bridges. Both the intersections and the bridges are represented by points in the feature classes or shapefiles. The samples used in the analysis would consist of all the points where the bridges are co-located with the road/water intersections (a small tolerance is allowed to account for round-off errors or mapping errors).
2. At least one data type is point data, the rest may contain polygons.
   An example of this would be coffee shops (points) vs. population density (polygons). In this case, the population polygons would be sampled at the point data locations.
3. All data types are polygons.
   A grid sampling scheme as described above is used.

It should be noted that *Aspect* currently does not handle the sampling of a sparse polygon layer. For sparse polygons, it is possible that none of the sampling locations will intersect the sparse polygons; thus, no data will be gathered. A workaround exists if the data is preprocessed to fill in the sparse data layer. The fill polygons should be given an attribute value to indicate that the sparse feature does not exist at that location. The intersect will then return a value that correctly shows the absence of a feature, rather than not returning any value at all.

9.2.3.2 Horizontal Rule Sampling

Several different types of sampling are used for the horizontal rules. For the distance relationship, buffers around the features are used to find the distances. The feature itself is clipped out of the buffer, resulting in a ring buffer. For varying distance ranges, multiple ring buffers are found by setting a larger buffer corresponding to the maximum distance range, and subtracting out a smaller buffer corresponding to the minimum distance range. The resulting ring buffer is used to intersect the second variable in the association. For the adjacent relationship, a small buffer is placed around one set of features, and sampling points are placed at equal intervals along the boundary of the buffer. The second feature set is sampled at those points.

9.2.4 Identification of Significant Variables

One major requirement of the vertical association rule analysis is the identification of the relevant variables to include in the rule. In this case, there is one “primary”

variable to predict, and there can be many "secondary" variables as predictors. Many data sets include a large number of candidate secondary variables, but not all of these variables will be strongly associated with the primary variable. Including too many variables will weaken the rules and distort the true relationship. Including too few variables may miss a significant relationship and could result in an inaccurate rule. If the variables to be examined for membership in an association rule can be determined in advance, it will reduce the computational effort required for the association rule discovery. Otherwise it is too time-consuming to exhaustively test every possible combination of variables in a rule.

Several methods are used together within *Aspect* to identify the relevant variables to include in a rule. Contingency table analysis (Chi Square) is used to find a measure of correlation between categorical variables. It can only be used between two variables, and does not provide information about cross-correlations, multivariate relationships, or nonlinear relationships. However, by performing this analysis between each pair of candidate variables, it can provide guidance for identifying which secondary variables should be included in the association with a primary variable. The *Apriori* algorithm (Agrawal and Srikant 1994) is a method used in association rules to find which combinations of variables have high support in a data set. It is assumed that if certain combinations of variables appear frequently, then they are the most relevant variables to include in an association rule. These two methods are discussed in detail here.

A final method to keep the variables in a rule relevant and interesting is to make use of the user's knowledge of the domain. Expert knowledge is used within *Aspect* in the following ways: through the selection of attribute names to include in the analysis (thus impacting the hierarchical variable tree), by binning the attribute values into fewer, more relevant categories, through the selection of primary/secondary variable candidates, and by utilizing the information gathered by the contingency table and *Apriori* analyses.

9.2.4.1 Contingency Table Analysis

Aspect allows the user to perform a contingency table analysis on candidate primary/secondary variable pairs, to see if the relationship between the two is strong. The contingency table analysis shows whether any two variables are related with statistical significance greater than that expected if they are randomly distributed. The results should be used with caution, since they do not include multivariate or nonlinear effects. However, they do provide some guidance, and in many situations it can be safe to either retain or eliminate a variable in the association rule based on the results from the contingency table.

Contingency table analysis is used to examine the relationship between two categorical variables, for vertical (overlapping) variables only. The values for each variable can be nonnumeric categories; if the variables are numeric, the numeric values are grouped together into bins. The table rows correspond to the categories or bins of the "primary" (or "dependent") variable, and the table columns correspond to the categories or bins of the "secondary" (or "independent") variable. Each cell in the table holds the number of times (the "frequency") that the two

Table 9.1 An example contingency table showing vegetation types vs. soil types. Each table cell shows the number of observed samples corresponding to each row and column

	Loam	Sand	Clay
Grassland	137	32	40
Woodland	56	6	2
Shrub	19	5	1

categories occur together in the data set. (The frequency is determined using a sampling scheme that depends on the data type; see Vertical Rule Sampling discussion above.)

An example of a contingency table is shown in Table 9.1. Suppose we are examining two variables, "Soil Type" and "Vegetation Type." The contingency table might look like this:

The numbers in each cell represent the samples in the data set. For example, the cell that is the intersection of "Loam" and "Grassland" shows 137 points. That means that the soil type "Loam" is sampled with the vegetation type "Grassland" 137 times.

Two tables are constructed. The first has the actual measured frequencies in the data set. These data are called "observed" data. The second table contains the expected frequencies for each cell that would occur if the two variables were independent of each other. These data are called "expected" data. The contingency table is used to compute a statistic called Chi Square. Chi Square is a measure of how different the observed data are from the expected data. If the Chi Square value is large (dependent upon the table's degrees of freedom and the allowable error), then the trend seen in the table is determined to be statistically different from the distribution that would occur from a random sampling, and can be assumed to be valid. This allows the computation of the correlation coefficient between the two variables and also the proportion of explained variance in the data.

A contingency table can have additional information included in each cell. By normalizing the frequencies by the total for each column, the result is the confidence for the two-variable association represented by the table cell (the table row corresponds to primary variable/consequent and the table column corresponds to the secondary variable/antecedent). The frequency itself is the support value (without normalization). The lift can also be computed. With this extra information, each table cell provides the strength of the two-variable association rule. The contingency table in Table 9.1 is expanded in Table 9.2 to include these extra values. Each cell contains the frequency, the confidence, and the lift.

The uppermost right cell in Table 9.2 shows the association rule for Clay vs. Grassland.

$$\text{Clay} => \text{Grassland}$$
$$\text{Support} = 40,\ \text{Confidence} = 93\%\ ,\ \text{Lift} = 1.33$$

Table 9.2 An expanded contingency table for vegetation type vs. soil type. Each table cell displays the number of samples, the confidence (in parentheses), and the lift (in parentheses)

	Loam	Sand	Clay
Grassland	137 (65%) (0.92)	32 (74%) (1.06)	40 (93%) (1.33)
Woodland	56 (26%) (1.23)	6 (14%) (0.65)	2 (5%) (0.22)
Shrub	19 (9%) (1.07)	5 (12%) (1.39)	1 (2%) (0.28)

Although the support is not as high as the Loam => Grassland rule, the confidence and lift make this a much stronger rule. In fact, since the lift for Loam => Grassland is less than 1.0, that rule is negatively correlated (i.e., the presence of loam is a negative indication of the presence of grassland).

The value of the contingency table is threefold. First, it provides guidance on which of many secondary variables has the greatest overall effect on the primary variable. Second, it shows which individual variable value pairs have a strong association. If a strong association is found in one table cell, such as Clay => Grassland above, then it is desirable to retain those variables in the association rule analysis, regardless of the overall strength of the table relationship. It is of particular importance to make note of these relationships, since the *Apriori* algorithm might miss this association if the support threshold for *Apriori* is set higher than 40.

A third advantage is that the contingency table forces the user to bin the data into categories resulting in a sample frequency sufficiently high for a valid Chi Square calculation. The analysis is valid only if no more than 5% of the cells in the expected table have frequencies less than or equal to five, and no cell frequency in the expected table is less than one. If the table frequencies are too low, attribute values can be binned together until the frequency criteria is met. The binning is useful also for association rule analysis, which also relies on categories. Any category with sufficient samples for the contingency table will also have sufficient samples for an association rule.

9.2.4.2 Illustration

This example (Table 9.3) shows how a contingency table can be interpreted in support of an association rule analysis. This table shows annual average mean temperature vs. vegetation/land use for Sonoma County, California. The three numbers in each cell are the sampling frequency (support), normalized frequency (confidence), and lift. One table cell in particular stands out, the one for Redwood vs. a mean temperature of 52˚– 55˚F (the coolest conditions). The lift value is 4.77. Lift values must be greater than 1.0 for the association to be true, and values greater

Table 9.3 Contingency table for Annual Average Mean Temperature vs. Land Use Categories in Sonoma County, CA. Each table cell displays the number of samples, the confidence (in parentheses) and the lift (in parentheses)

Observed values	Mean Temp 52–55 [˚F]	Mean Temp 55–57 [˚F]	Mean Temp 57-59 [˚F]	Mean Temp 59-61 [˚F]
Grassland	33	62	220	50
	(19.5%)	(20.9%)	(22.6%)	(17.8%)
	(0.92)	(0.98)	(1.07)	(0.84)
Woodland	15	30	73	24
	(8.9%)	(10.1%)	(7.5%)	(8.5%)
	(1.07)	(1.22)	(0.91)	(1.03)
Redwood	84	56	37	2
	(49.7%)	(18.9%)	(3.8%)	(0.7%)
	(4.77)	(1.81)	(0.37)	(0.07)
Montane Hardwood	15	63	271	84
	(8.9%)	(21.2%)	(27.9%)	(29.9%)
	(0.35)	(0.84)	(1.11)	(1.19)
Other	5	2	21	15
(Barren/Wetlands)	(3.0%)	(0.7%)	(2.2%)	(5.3%)
	(1.18)	(0.27)	(0.86)	(2.13)
Douglas Fir	7	46	50	13
	(4.1%)	(15.5%)	(5.1%)	(4.6%)
	(0.61)	(2.30)	(0.76)	(0.69)
Crops	2	27	168	65
	(1.2%)	(9.1%)	(17.3%)	(23.1%)
	(0.08)	(0.60)	(1.13)	(1.52)
Shrub	6	11	43	18
	(3.6%)	(3.7%)	(4.4%)	(6.4%)
	(0.78)	(0.82)	(0.97)	(1.41)
Urban	2	0	89	10
	(1.2%)	(0%)	(9.2%)	(3.6%)
	(0.20)	(0.00)	(1.56)	(0.61)

than 2.0 tend to be rare. A value of 4 or 5 is an exceptionally strong association. The support for that cell is 84. If *Apriori* is run with a support threshold higher than this (to increase computational efficiency), then it is important to manually include the mean temperature vs. Redwoods candidate variables in the association rules. The Barren/Wetlands categories were combined into an "Other" category, since each had insufficient points to remain separate for a valid Chi Square calculation or to be regarded in an association rule.

9.2.4.3 Apriori Algorithm

Aspect incorporates a known algorithm for identifying frequent itemsets in a vertical association data set. This algorithm, called *Apriori* (Agrawal and Srikant 1994), finds frequent combinations of variables that meet a user-defined support threshold. The combinations can include any number of variables up to the maximum contained in the data set. It begins by finding frequent one-itemsets, then two-itemsets, up to k-itemsets, where k is the maximum number of variables. After finding the

frequent n-itemsets, this information is used to prune the n+1-itemset. Since those rule sets that contain infrequent itemsets are pruned early, the association rule discovery is much faster than with an algorithm that examines every rule combination. *Apriori* runs fast for high support thresholds and smaller numbers of variables. As the number of variables increases and/or the support threshold decreases, *Apriori* slows down considerably.

There are several disadvantages to the *Apriori* algorithm.

1. It is difficult to filter the output
   Apriori can generate a great deal of output, especially for large numbers of variables and a low support threshold. For example, if a four-variable combination is found, the algorithm will also output all the two-variable and three-variable subsets of it. The lesser combinations might not be of interest, and they clutter the output. However, the two- and three-variable combinations are of interest to the user in some cases, in particular if the confidence of the lower-variable combination is much higher than the confidence of the higher-variable combination. Also, some two- or three-variable combinations are output that are not subsets of a four-variable combination, so they are of interest themselves.
2. Results are based on the support threshold only
   Once combinations based on support are found, a second pass through the data is required to compute the confidence and lift. (It is necessary to have the primary and secondary variables specified to compute the confidence and lift.) *Apriori* can easily miss strong associations that have high confidence and lift but have a support value below the threshold. The temptation when running *Apriori* is to keep the support threshold high, both to limit the amount of output and to control the computational time. The user naturally wants to find only those combinations of variables that have high support for a rule. However, as was seen in the contingency table example, the strongest associations in terms of confidence and lift might not necessarily have the highest support values. Thus, *Apriori* works best in conjunction with contingency table analysis to identify the strongest associations.

In *Aspect*, the association rule analysis with the *Apriori* algorithm completes two passes through the data. The first pass finds the variable combinations that meet or exceed the support threshold, and the second pass computes the confidence and lift for those combinations. Only those combinations that meet all three thresholds are output to the user.

9.3 Results

9.3.1 Land Use Discrimination (Vertical Associations)

This analysis illustrates the use of multiple data layers, the selection of relevant variables, and attribute value binning. In this example, the vertical association rule

analysis was applied to the discrimination of vegetation or land cover types based on climate, soil, and socioeconomic features. Land cover maps are typically derived from remote sensing images, and association rules could provide additional information for the identification and interpretation of the various classes. The following analysis was performed for Sonoma County, California.

The land cover dataset used in this analysis is the USDA CALVEG dataset (USDA Forest Service RSL 2003), which uses the California Wildlife Habitat Relationships (WHR) classification system. This dataset is derived from Landsat TM images at 30 m resolution with a minimum mapping unit of 1 ha. Soil data is from the USDA SSURGO project, socioeconomic data from the 2000 US Census at the block group level, and climate data covering the years 1961–1990 from PRISM group at Oregon State University (Daly et al. 2001). Both the numerical and the categorical data are binned to create a smaller number of categories, so that the frequency of data for each category is high enough for the contingency table analysis and for the association rule support. For example, twenty-two land cover categories in the county were combined into nine categories (urban, Redwood, montane hardwood, Douglas Fir, crops, grassland, shrub, other woodland, barren/wetlands).

Soil categories were combined into nine groups including loam, clay, silt, sand, and soil type derivatives. Socioeconomic data are the population densities for all residents, those under age five, and those over age 65, the household density, the average household size, the owner-occupied density, the renter-occupied density, and the median household income. Road density was included as before. Climate data included annual average maximum, minimum, and mean temperature, annual average precipitation, annual average relative humidity, average number of days above 90˚F, and average number of days below 32˚F. Terrain data was provided as elevation values. The climate and elevation data were all originally in a raster format. The climate data was at approximately 2 km resolution, and the elevation data was at 30 m resolution resampled to 2 km to match the climate data. These raster data sets were vectorized into polygon shapes to use in the analysis.

The contingency table analysis showed that all variables were statistically significant with respect to land cover, with a shared variance ranging from 3 to 25%. By inspecting the individual cells of the contingency tables, it was determined that the median age and the number of days under 32˚F could be safely eliminated from the association rules due to low shared variance and low support and confidence for each combination. Although median household income and number of days above 90˚F also had lower shared variance, it was determined that these variables should be retained since individual values showed high support and confidence with individual land cover categories. All other variables were retained for consideration in the *Apriori* algorithm.

The *Apriori* algorithm was used to generate the frequent itemsets of the retained variables, and the support, confidence, and lift of each combination was computed. Below are listed some of the rules for the “Redwood” land use class. These results illustrate the tradeoff between the support and confidence thresholds. The population and household densities are provided as number per hectare.

MaxTemp(60°−65°F)^MinTemp(42°−45°F)^MeanTemp(52°−55°F)
=> Redwood
Support = 80 (4.6%), Confidence = 60%, Lift = 5.9

MaxTemp(60°−65°F)^MinTemp(42°−45°F)^MeanTemp(52°−55°F)
^Household Density(0 − 0.1) => Redwood
Support = 80 (4.6%), Confidence = 62.5%, Lift = 6.1

MaxTemp(60°−65°F)^MinTemp(42°−45°F)^MeanTemp(52°−55°F)
^Elevation(500' − 1000') => Redwood
Support = 42 (2.4%), Confidence = 72%, Lift = 7.1

These three rules show that cooler temperatures are a good indicator of redwood locations, or conversely, that redwood forests lower the temperature. (The range of annual average maximum temperature for the entire county is 60°–75°, the range for the minimum temperature is 40°–50°, and the range for the mean is 50°–61°.) Including the low household density increases the strength of the rule without reducing the support. Including the elevation increases the confidence and lift even more, but at the expense of the support. Including both the household density and the elevation gives the same numbers as the elevation, showing that the household density holds throughout that elevation range:

MaxTemp(60°−65°F)^MinTemp(42°−45°F)^MeanTemp(52°−55°F)
^Household Density(0 − 0.1)^Elevation(500'−1000') => Redwood
Support = 42 (2.4%), Confidence = 72%, Lift = 7.1

The temperatures combined with the precipitation and soil type further increase the confidence and lift at the expense of support. The confidence is increased to 100%, which means for data with those values of temperature, precipitation, and soil type, the land use class is always Redwood. However, very few points meet all those requirements; thus the support is low.

MaxTemp(60°−65°F)^MinTemp(42°−45°F)^MeanTemp(52°−55°F)
^Precipitation(37"−45")^SoilType("Hugo Loam, 30% − 50% slopes")
= > Redwood
Support = 9 (0.5%), Confidence = 100% , Lift = 9.7

Examples of the multivariate association rules for crops are as follows:

Elevation(0'−100') = > Crops
Support = 104 (6%), Confidence = 48% , Lift = 3.2

Elevation(0'−100')^MeanTemp(57°−59°F)
^AvgHouseholdSize(2.5 − 3.0)^PopulationDensity(0 −1.0)
^SoilType("Reyes Silty Clay, 0% − 2% slopes") = > Crops
Support = 19(1%), Confidence = 49%, Lift = 3.2

The probability that cropland is present can be predicted with elevation alone. Adding more variables does not change the confidence and lift, but reduces the support.

9.3.2 Road Density vs. Urban Areas (Horizontal Associations)

The computation of a higher order quantity such as density occurs in horizontal association analysis. This is a significant example, since the interpretation of remote sensing images can be enhanced by the knowledge of ground information. One key piece of information is the existence of urban boundaries. If an urban area is known, this could help to interpret whether a pixel represents asphalt or bare dirt, for example. Thus, a relationship that determines urban areas from well-defined features in a remote sensing image would be useful. Road densities can be easily measured; thus, an association rule that relates road density to urban areas is sought.

This analysis was performed using data from Sonoma County in northern California (Fig. 9.1). Road and urban area data layers were obtained from the US Census. Sonoma County has well-defined urban areas surrounded by green belts and agricultural regions. The county extent is approximately 80 km north-south and 100 km east-west. The geographic domain was divided into 60×60 grid cells, each approximately 1.7×1.35 km. Within each grid cell, the road density was computed as the total length of all road features per grid cell area. A random point was chosen within each grid cell for sampling the urban layer. Of the 3600 grid cells, 1779 of them intersect the county shape. The road densities range from 0 to 36,000 m/grid cell area.

An initial analysis was performed to get a concept of the density values in the data set and the distribution of the associations. Examining these results (in particular the confidence), it is seen that the road density is strongly associated with "Not

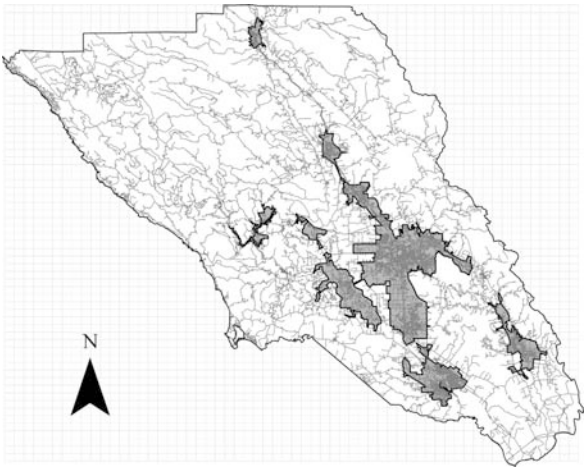

Fig. 9.1 The road network in Sonoma County (*lines*) and the urban areas (*polygons*). The *grid cells* show the sampling resolution

Table 9.4 The final results for the road density vs. urban areas

Association rule	Support	Confidence (%)	Lift
Density(roads, <12,500) => feature(not urban)	1604	95	1.05
Density(roads, >12,500) => feature(not urban)	17	17	0.19
Density(roads, <12,500) => feature(urban)	77	5	0.52
Density(roads, >12,500) => feature(urban)	81	83	9.31

Urban" for densities under 12,000 (approximately), and the road density is strongly associated with "Urban" for densities above 12,000 (approximately). With further experimentation, it is seen that a density of 12,500 is a good density value to use to discriminate between urban and non-urban areas (Table 9.4).

It is interesting to note that the lift in this example does not agree well with the confidence. The lift is defined as the confidence normalized by the probability of the consequent occurring. Examining the rules in Table 9.4, the lift is seen to be low for "Not Urban" rules, and high for "Urban" rules. This is because there are few urban data points in this data set compared with non-urban data points. Thus, the probability of the consequent occurring is low for urban points (increasing the lift), and the probability of non-urban points occurring is high (decreasing the lift). This unbalanced distribution of data points may be skewing the lift calculation.

9.3.3 Water Features vs. Adjacent Soil Type (Horizontal Associations)

This example illustrates how to create and analyze a horizontal rule using the adjacent predicate. This example seeks to identify which soil types occur frequently adjacent to a river bank. The study area is the area surrounding the Russian River in Sonoma County, California (Fig. 9.2). The soil data set is taken from the USDA SSURGO project. The soil type attribute in this feature class includes a code for water bodies; thus, only one data layer is used in the analysis. The resulting rule should show which soil types are most common along the river. The rule template is

$$\text{Feature\_At(X, Water)} => \text{Feature\_At(Y, Soil(Soil Type, value))} \wedge \text{Adjacent(X,Y)}$$

A buffer distance of 10 m was used to sample the soil polygons adjacent to the river. Since the river and the soil data come from the same feature class, there is no concern with this data set about boundaries not matching. The sampling points are spaced approximately 1 km apart (Fig. 9.3). The analysis was run with a support threshold of 10 and a confidence threshold of 5. Experimentation showed that these values eliminate the rules with a support of only one or two.

Fig. 9.2 The soil layer for the Russian River area in Sonoma County, California

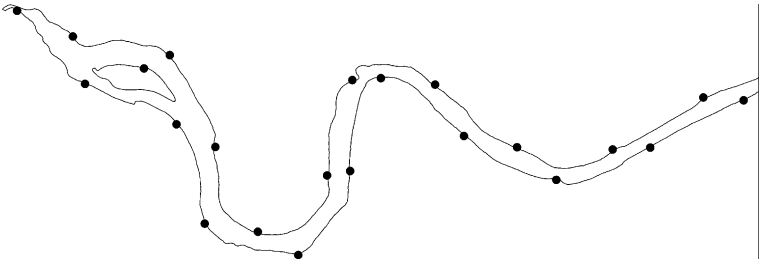

Fig. 9.3 The points adjacent to the Russian River used for sampling. Only a portion of the river is shown. The sample points are spaced approximately 1 km apart

The association rule results are shown in Table 9.5. Five soil types appear with a fairly high frequency: AdA (Alluvial Land, Sand), HkG (Hugo Very Gravelly Loam, 50–75% slopes), RnA (Riverwash), YlA (Yolo Sandy Loam, 0–2%), and YmB (Yolo Sandy Loam, Overwash, 0–5%).

9.4 Conclusions

A data mining system, called *Aspect*, was designed and developed to allow GIS users to perform association rule analyses in geographic data. *Aspect* reads in standard data formats and can be used as an extension to commercial GIS software. One of the difficulties in association rule data mining is the identification of relevant variables to include in a rule, so that accurate and interesting rules are found without undue computational burden. *Aspect* provides multiple methods to provide guidance

Table 9.5 The association rules showing soil types adjacent to a river

Association rule	Support	Confidence
FeatureAt(X: Water)) => FeatureAt(Y: Soil Type, AdA)) ^ Adjacent(X,Y)	28	
FeatureAt(X: Water)) => FeatureAt(Y: Soil Type, HkG)) ^ Adjacent(X,Y)	12	
FeatureAt(X: Water)) => FeatureAt(Y: Soil Type, RnA)) ^ Adjacent(X,Y)	23	
FeatureAt(X: Water)) => FeatureAt(Y: Soil Type, YlA)) ^ Adjacent(X,Y)	11	
FeatureAt(X: Water)) => FeatureAt(Y: Soil Type, YmB)) ^ Adjacent(X,Y)	19	

AdA Alluvial Land, Sandy; *HkG* Hugo Very Gravelly Loam, 50–75% slopes; *RnA* Riverwash; *YlA* Yolo Sandy Loam, 0–2% slopes; *YmB* Yolo Sandy Loam, Overwash, 0–5% slopes.

for variable selection, and the user interface allows the user to iterate between these methods to narrow down the rule candidates, based on the support, confidence, and lift thresholds and the number of desired rules.

This system has been demonstrated to handle high dimensional data sets and to find the appropriate variable combinations that form strong rules. These results can be used as an alternative to multivariate statistical analysis, which can be difficult for users without the appropriate statistical expertise to perform correctly, in particular for categorical data types. Association rules are applicable to the interpretation of remote sensing images.

Acknowledgments This work was sponsored by the U. S. Army Topographic Engineering Center, Vicksburg Consolidated Contracting Office, under Contract No. W9132V-04-C-0025. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the US Army Corps of Engineers.

Robert Huizar III, of the University of North Carolina at Charlotte, assisted with the Sonoma County data sets. His contributions are gratefully acknowledged.

References

Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI (1996) Fast Discovery of Association Rules. In: Fayyad UM et al. (eds) Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, CA, pp. 307–328

Agrawal R, Srikant R (1994) Fast Algorithms for Mining Association Rules. In: Proc. 1994 Int. Conf. Very Large Data Bases. Santiago, Chile, pp. 487–499

Daly C, Taylor GH, Gibson WP, Parzybok TW, Johnson GL, Pasteris, P (2001) High-Quality Spatial Climate Data Sets for the United States and Beyond. Trans. Am. Soc. Agric. Eng. 43:1957–1962

Han J, Kamber M (2001) Data Mining: Concepts and Techniques. Academic Press, San Diego, p 261

Hipp J, Guntzer U, Nakhaeizadeh G (2000) Algorithms for Association Rule Mining: A General Survey and Comparison. In: SIGKDD Explorations. 2:58–64

King RB (2002) Land Cover Mapping Principles: a Return to Interpretation Fundamentals. Int. J. Remote Sens. 23:3525–3545

Koperski K, Han J (1995) Discovery of Spatial Association Rules in Geographic Information Databases. In: Proc. 4th Int. Symp. on Large Spatial Databases, Maine, pp 47–66

Rodman LC, Jackson J (2007) Spatial Association Rule Discovery with Rule Classification and Variable Sensitivity. NEAR TR 624, Nielsen Engineering & Research, Mountain View, CA

Scheffer T (2001) Finding Association Rules that Trade Support Optimally Against Confidence. In: Lecture Notes in Computer Science, Springer, 2168:424–435

USDA Forest Service RSL (2003) CALVEG Vegetation Mapping Program, 1920 20th Street, Sacramento, CA 95814

USDA SSURGO, Natural Resources Conservation Service, Soil Survey Geographic (SSURGO) Database for Sonoma County, CA

WHR Classification, CA Department of Fish and Game, Biogeographic Data Branch

Chapter 10
An Intelligent Archive Testbed Incorporating Data Mining

H.K. Ramapriyan, D. Isaac, W. Yang, B. Bonnlander, and D. Danks

10.1 Introduction

Many significant advances have occurred during the last two decades in remote sensing instrumentation, computation, storage, and communication technology. A series of Earth observing satellites have been launched by several countries around the world and have been operating and collecting global data on a regular basis. These advances have created a data rich environment for scientific research and applications. NASA's Earth Observing System (EOS) Data and Information System (EOSDIS) has been operational since August 1994 with support for pre-EOS data. Currently, EOSDIS supports all the EOS missions including Terra (launched in 1999), Aqua (launched in 2002), ICESat (launched in 2002) and Aura (launched in 2004). EOSDIS has been effectively capturing, processing and archiving several terabytes of standard data products each day. It has also been distributing these data products at a rate of several terabytes per day to a diverse and globally distributed user community (Ramapriyan et al. 2009). There are other NASA-sponsored data system activities including measurement-based systems such as the Ocean Data Processing System and the Precipitation Processing system, and several projects under the Research, Education and Applications Solutions Network (REASoN), Making Earth Science Data Records for Use in Research Environments (MEaSUREs), and the Advancing Collaborative Connections for Earth-Sun System Science (ACCESS) programs. Together, these activities provide a rich set of

H.K. Ramapriyan (✉)
NASA Goddard Space Flight Center, Greenbelt, MD, USA
e-mail: rama.ramapriyan@nasa.gov

This work was performed by the first author as part of his official duties as an employee of the US government. It was supported by the NASA's Science Mission Directorate. The remaining authors were supported under Cooperative Agreement NCC5-645 between NASA and George Mason University. The opinions expressed are those of the authors and do not necessarily reflect the official position of NASA.

L. Di, H.K. Ramapriyan (eds.), *Standard-Based Data and Information Systems for Earth Observation*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-540-88264-0_10,

resources constituting a "value chain" for users to obtain data at various levels ranging from raw radiances to interdisciplinary model outputs. The result has been a significant leap in our understanding of the Earth systems that all humans depend on for their enjoyment, livelihood, and survival.

The trend in the community today is towards many distributed sets of providers of data and services. Despite this, visions for the future include users' being able to locate, fuse and utilize data with location transparency and high degree of interoperability, and being able to convert data to information and usable knowledge in an efficient, convenient manner, aided significantly by automation (Ramapriyan et al. 2004, NASA 2005). We can look upon the distributed provider environment with capabilities to convert data to information and to knowledge as an Intelligent Archive in the Context of a Knowledge Building system (IA-KBS). Some of the key capabilities of an IA-KBS are: Virtual Product Generation, Significant Event Detection, Automated Data Quality Assessment, Large-Scale Data Mining, Dynamic Feedback Loop, and Data Discovery and Efficient Requesting (Ramapriyan et al. 2004).

Large-Scale Data Mining (LSDM) plays a very important role in IA-KBS. Two important uses of LSDM are: retrospective studies covering large temporal and geographic extent; and precursor detection, where indicators of significant events are identified through analysis of historical data. Note that, once good precursors have been identified via a scientific LSDM process, computationally efficient filters on real-time data streams can be constructed so that significant events can be detected in observational data in near real-time and users can be alerted.

Especially relevant to data mining in the above discussion, there have been several research investigations in the area of Intelligent Data Understanding that were supported by NASA's Intelligent Systems Project under the Computing, Information and Communication Technology Program that can contribute to the goals of an IA-KBS. However, these investigations typically perform proofs of concept on a relatively small scale. Before their contributions can be implemented on a large scale commensurate with today's Earth science data archives, it is necessary to test them in a pseudo-operational environment. In this chapter, we describe the implementation of a testbed to accomplish this and discuss some of the observations and lessons learned from its implementation. This is a more detailed discussion than the summary of the implementation and results presented in Ramapriyan et al. (2005).

In Sect. 10.2, we present the basic concepts of an IA-KBS. In Sect. 10.3, we discuss the goals of the testbed and describe the application scenario (prediction of wild fire potential from historical and current remote sensing observations) being tested in the testbed. In Sect. 10.4, we provide a description of the algorithm used and implementation details. In Sect. 10.5, we present the results of implementation including derived fire prediction maps, processing speeds and feasibility of pseudo-operational implementation. In Sect. 10.6, we document our conclusions and lessons learned.

10.2 Intelligent Archives

The concept of intelligent archives was spawned out of long experience with Earth science data archives and recognition of the convergence of two factors: the accumulation of enormous quantities of valuable scientific data and the increasing practicality of applying machine learning techniques to large-scale data sets. The result was a growing belief that current data archives could and should evolve to better support the scientific process and help unlock the untapped potential of both current data holdings and ongoing observational data streams. We believe the time has come for intelligent archives, which not only store and disseminate data, but also play an active role in the knowledge building process.

Today's data archives play a key role supporting the value chain that turns data into information and knowledge. They provide common repositories for the collection and dissemination of information at all levels, from raw observations to high-level analyses. And yet it is clear that they have the potential to provide much more value in the overall value chain. For example, archives are uniquely positioned to perform the following useful functions:

- Mining archived data holdings to add metadata and thereby improve data access and usability;
- Identifying quality issues as data are being stored, while there is still the opportunity to recapture the affected observations;
- Detecting unusual patterns in data that may indicate an event of interest; and
- Facilitating collaboration and exchange among distributed research groups.

An assessment of the potential role and function of an intelligent archive identified six key capabilities such a system should exhibit: virtual product generation, significant event detection, automated data quality assessment, large-scale data mining, dynamic feedback, and data discovery and efficient requesting.

10.2.1 Virtual Product Generation

An archive does not need to produce and archive all of the derived data products that will be requested of it in advance. In many cases, only a small percentage of the products generated are actually requested. Virtual product generation allows a user to treat a product as though it were being retrieved from the archive when, in reality, the data inputs are automatically retrieved, assembled, and processed into the desired form "on the fly," in response to the request (Clausen and Lynnes 2003). This adds latency but can result in significant storage cost savings and eliminate the need for reprocessing as algorithms are improved. An intelligent archive minimizes latency by anticipating demand (e.g., based on predictive models of usage patterns and significant event detection), computing needed products just in time with a relatively small percentage of the total data.

Another aspect of virtual product generation is the ability to assemble, transparently, inputs to the production algorithm from a variety of sources and locations. This requires a global registry, interface standards, and supporting middleware (so that production algorithms receive the data in an acceptable and consistent format).

Further, a virtual product generation capability should optimize the assignment of processing resources by minimizing the need for data communications, and considering current resource utilization and availability. This means building and using a global predictive model of the interconnected network of storage and processing resources, and keeping the model current via frequent state updates (where "frequent" might mean on the order of tens of seconds).

10.2.2 Significant Event Detection

With a constant stream of several terabytes of data per day entering an archive, it is likely that manual analyses will miss some significant events, or at least miss them until the opportunity to perform focused collection of additional information related to the event has passed.

Significant event detection helps identify phenomena of interest within very large data sets and data streams. This in turn enables not only near-real-time reporting of geophysical events in an ingest data stream, but also content-based queries if the events are stored as metadata. The idea here is to place matched filters or pattern recognition algorithms on an input data stream (or streams) to automatically detect the occurrence of an event. Some examples of significant events are hurricanes, wild fires, volcanic eruptions, and failure of a sensor or some other part of the information processing chain. The event detection, in turn, could trigger a variety of actions such as notification of subscribers, generation and distribution of associated products, retasking of sensor assets, reallocation of system resources, and self-repair.

10.2.3 Automated Data Quality Assessment

As in the case of significant events, experience has shown that data quality issues can stay undiscovered for long periods, hidden in the flood of data. In addition to surfacing such issues in a timely manner, the potential exists to circumvent a variety of complex data quality issues.

Automated data quality assessment maintains the algorithmic processing pedigree of a data product, and ensures the scientific and algorithmic consistency of the underlying modeling and processing assumptions (Isaac and Lynnes 2003). An intelligent archive can take responsibility (to a greater or lesser extent) for monitoring and perhaps correcting the quality of the products it delivers to a requestor or a consuming process. This includes both the algorithms and inputs used to generate the product (that is, a sophisticated kind of product provenance and configuration

control) as well as internal inspection of the products to ensure that they meet a variety of user-specified characteristics (e.g., regarding cloud cover, dynamic range, or sampling resolution). Other sources of error (e.g., bit errors, compression/decompression lossiness and mistakes in indexes or metadata) can also be detected prior to delivery and, in some cases, corrected or ameliorated (e.g., through interpolation or other data modeling approaches).

10.2.4 Large Scale Data Mining

While data mining has already proven to have utility in Earth science data analysis, an intelligent archive must perform data mining efficiently to enable the analysis of large data volumes. Here, the primary meaning of the term is a process that finds higher-level emergent causal relationships at a modeling level above the level at which the inputs exist. Typically, data mining sits at the top of the value chain – taking information as input, and producing knowledge. Two important examples of this type of analysis are (1) retrospective studies covering large temporal and geographic extent, and (2) precursor detection, where indicators of significant events are identified through analysis of historical data. Note that once good precursors have been identified via this scientific data mining process, computationally efficient filters on real-time data streams can be constructed. The output of a successful data mining process is typically a model that can serve as the basis for prediction, event detection, classification, quality assessment, or other purposes. The knowledge derived from successful data mining may also have other value or utility: pure science (discovering or confirming previously unverified correlations or relationships – e.g., in global climatology); efficiency (discovering that only some inputs contribute significantly to an output); and instrument or spacecraft health and safety (detection or prediction of anomalies).

10.2.5 Dynamic Feed-back

The ability to stochastically optimize the allocation of the storage, network, and computing resources of the archive using dynamic feed-back loops supports the other archive functions by increasing throughput and reducing user-experienced latency in product delivery (Morse et al. 2003). An intelligent archive can be modeled (McConaughy and McDonald 2003) as one component in a complete system that includes sensor tasking, sensors and collection, ground station early level product generation, archiving and higher-level production, real-time or near real-time applications, and user request satisfaction and associated production. There are two low-latency feedback loops of interest. The first is for retasking of resources to support time critical scenarios (e.g., fire detection). The second is for system resource optimization to maximize throughput and minimize latency of delivered user-requested products.

10.2.6 Data Discovery and Efficient Requesting

Some advanced capabilities of an intelligent archive, particularly virtual product generation, have the potential to put heavy loads on cooperating systems. For that reason, an intelligent archive should act as an intelligent requestor of data, exploiting its knowledge of information interrelationships and computing resources to minimize its load on cooperating systems. In addition, intelligent archives of the future should be able to detect the presence of newly available data anywhere in the world (Isaac and McConaughy 2004), determine the usefulness of the data, learn how to access them and ultimately provide the data to users or applications in a usable form. This involves an ongoing search process that keeps constantly and persistently aware of potential data sources and their changing status, and capabilities to retrieve and reformat the data transparently to meet the users' data interface requirements or preferences. In this way, the intelligent archive becomes an interface not only to its own holdings, but also to the broad array data sources of interest across the accessible web. Together or independently, these capabilities have the potential to significantly improve support of the knowledge-building value chain.

10.3 Testbed and Scenario

The concept of an intelligent archive is intriguing. But it is even more interesting if it is feasible to implement the concept using current computing technology. A test using a realistic scenario on realistic data volumes suggests that it is.

Perhaps the most important capability identified for an intelligent archive is the ability to facilitate the transformation of data into knowledge in a distributed environment using large-scale data mining. Data mining algorithms are generally viewed as unsuited for large-scale use disciplines like Earth science that involve very high data volumes. There have been many research projects that have developed algorithms with promising results. But they have generally been tested on data subsets of only a few gigabytes, while large-scale datasets tend to be multiple terabytes in size. To bridge the gap between research and operations, we constructed a testbed to see how a typical algorithm would perform on full-scale data sets.

The testbed provides the computational and data resources required for implementing IA-KBS concepts on a scale that provides concrete evidence about the associated benefits and risks prior to implementing these concepts in operational systems. Such evidence is important both to the users of information and to data managers. The testbed provides a capable and flexible infrastructure for exploring a variety of data mining scenarios, though the focus in this discussion is primarily on performance and utility outcomes rather than system and software components or frameworks.

Using the six key capabilities identified above as a guide, several NASA-funded research projects and their algorithms were surveyed and assessed for their applicability to the testbed. Initially, a "reference architecture" for the IA-KBS was prepared (Morse and Yang 2004). In this reference architecture, key interfaces and

dependencies were specified, overlapping or redundant functionality was eliminated, and sample use cases and associated operations' concepts were created and described. The reference architecture makes evident where a given research algorithm or capability might reside. The goal was to find research that shows both scientific and operational relevance, and that is applicable to as large a subset of the IA-KBS functional capability as possible. Other evaluation criteria that entered into the selection process included implementation feasibility, source data availability, and collaboration potential.

The problem selected for demonstration is fire prediction (Bonnlander 2005). Fire prediction is very relevant socially. Also, it is a challenging problem, since there is a large component of stochastic uncertainty. Fuel type and availability, moisture, sources of ignition, temperature and precipitation –over considerable lengths of time and seasonal conditions – influence the predicted fire potential. Analysis shows that the algorithm can exercise most of the functional areas identified in the IA-KBS reference architecture. Finally, the scientific goals of the research will benefit from the testbed, since the testbed can process a large variety of geographic areas, fuel types, and seasonal conditions, and hence significantly extend the scientific relevance of the algorithm into new and previously unexplored aspects of the underlying phenomena.

10.3.1 Design Issues

The IA-KBS project had identified a number of technical issues associated with implementation of the intelligent archive concepts, most notably scalability issues (McConaughy and McDonald 2003, Isaac and McConaughy 2004). These issues needed to be addressed in a manner that was operationally relevant and yet could be implemented on a limited budget.

- Scalability and Parallelization. Two approaches to parallelization were considered: coarse grained and fine grained. Coarse grained parallelism was chosen because, although it requires partitioning the data and generating multiple independent models, it is much easier to run multiple instances of a data mining algorithm on each node than it is to implement a truly distributed algorithm. In the fire prediction scenario, it was convenient to partition data along fuel type (land-cover) and month of the year.
- Source Data Restructuring. Most data mining algorithms require observational data with one record per observation having all the relevant independent variables, while remote sensing data are typically stored as files with parameters covering contiguous areas in space and time. The fire prediction algorithm was no exception to this. Although an indexing scheme could be used to map from one representation to another, the resulting physical data access would "bounce" around the source data, resulting in poorer performance. Therefore, we chose to physically re-order the data into the form and format expected by the data mining algorithms.

- Representation of Time. It is obvious that, since prior precipitation has a significant effect on fire potential, prediction of fires involves time explicitly. However, there is no explicit mechanism in the algorithms selected for the testbed for handling time. Instead, observations for the same variable at different times are simply transposed into multiple variables within a single record for input into the data mining algorithm. This temporal "flattening" is straightforward unless the time series is very long, in which case the dimensionality of the data space would grow too large. For the fire prediction scenario, we kept the time series short by reducing measurements into a few averages for the prior day, week, month, etc.

10.3.2 Testbed Design

Prior work in the IA-KBS project identified general capability needs, challenges, and opportunities (Ramapriyan et al. 2004). The testbed design provides an opportunity to explore these general concepts at a practical level that would be relevant to an operational system. The testbed design is discussed briefly below as three different views: a system network view, a functional view, and a software component view.

10.3.2.1 System Network View

One of the most important aspects of the testbed is that it should demonstrate intelligent archive concepts in an operationally-relevant environment without interfering with the production operations of an actual operational system. Features included to make the environment "operationally relevant" include a high-performance node for the data mining and event detection components, use of pre-production and production archive nodes for source data, and high-speed networks for node connections. Figure 10.1 shows a simplified view of the components from the system network's point of view. The peer archives shown here are two of several Distributed Active

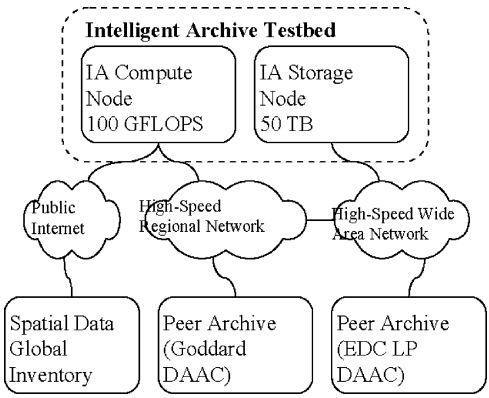

Fig. 10.1 Intelligent archive testbed system components

Archive Centers (DAACs) that operationally archive and distribute NASA's Earth science data. For the purposes of the testbed, data are obtained from these DAACs and stored in a separate IA storage node.

10.3.2.2 Functional View

The testbed includes a subset of the functional components identified in the IA-KBS reference architecture (Morse and Yang 2004), which were derived directly from the envisioned IA capabilities. The primary focus of the testbed is on the data mining and event detection components. These two components work together: the data mining component examines historical data to extract a statistical model of fire potential; the event detection component then uses this model to scan current data to assess current fire potential (Fig. 10.2).

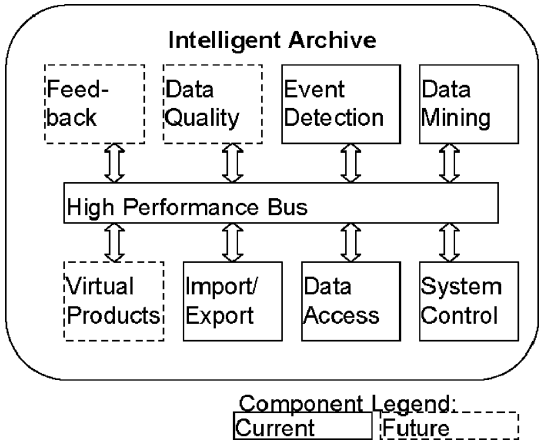

Fig. 10.2 Intelligent archive functional components

The data mining algorithms process the prepared data and extract a statistical model of fire potential based on the available remote sensing parameters. The algorithm implemented is logistic regression, since it performs well in terms of both accuracy and computational effort. The data mining algorithms are implemented in MATLAB.

10.3.2.3 Software Component View

The testbed includes a variety of software components that together provide the infrastructure needed to manipulate and mine large volumes of data. These are grouped as shown in Fig. 10.3 into layers of services at differing levels of abstraction.

The Local Application Platform Layer provides Job Management and Data Access services. Job Management includes the MATLAB Distributed Computing Toolbox/Engine for dispatching different data pre-processing, data mining, and

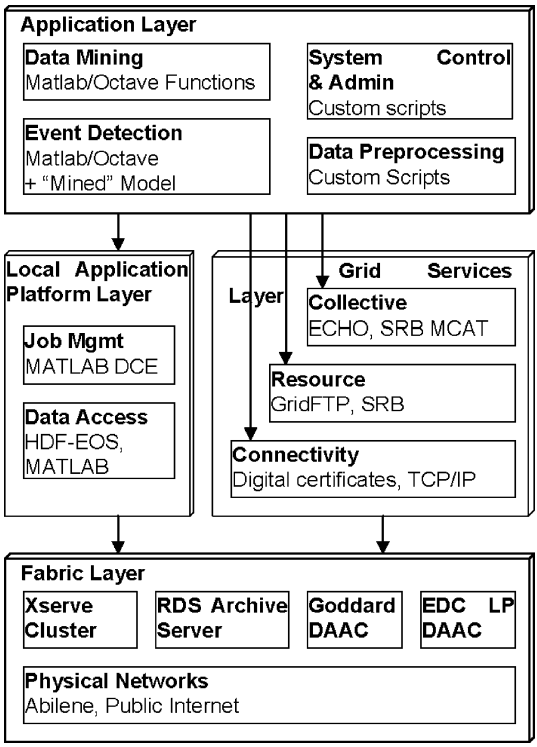

Fig. 10.3 Intelligent archive software components

event detection tasks to different IA Computational Node processors. Data Access services include Hierarchical Data Format (HDF) libraries for reading NASA remote sensing data files, and MATLAB I/O for managing pre-processed data.

The Grid Services Layer provides a variety of services for locating and accessing distributed computing and storage resources. The testbed uses these services mainly to identify and obtain source data for use by the data mining and event detection components in the Application Layer. The Collective services employ the EOS Clearinghouse (ECHO) for identifying specific files that contain the remote sensing parameters for the times and locations of interest. The Resource services are used primarily to access data from the grid using GridFTP. The Connectivity services include services for authenticating the local server to the grid, plus low-level communication services.

10.3.2.4 Data Preparation and Mining

As noted above, the data mining problem selected for demonstration of the IA testbed is wildfire prediction. Fire potential is determined by a number of parameters for which good remote sensing data exist, including temperature, precipitation history, fuel type and availability, and fuel moisture. Ignition events, including

lightning and human activities, are the final factor in the occurrence of wildfires. However, these are excluded from the predictive model because these are immediate causes of fires rather than predictors of fire potential several days ahead of time (Fig. 10.4).

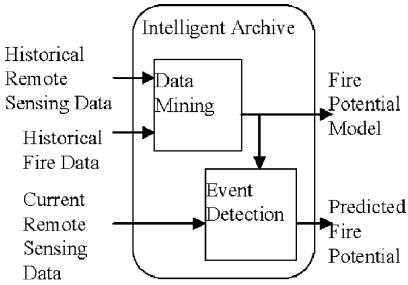

Fig. 10.4 Research scenario

10.4 Testbed Implementation

10.4.1 The Testbed Algorithm Implementation

Algorithm: The algorithm follows a set of steps that reflect a fairly standard approach to statistical forecasting. This approach involves defining a collection of independent variables (in this case, variables representing climate and historical fire occurrence data) and a single dependent variable (the occurrence of at least one fire within the next N days) in order to construct a model that produces a probability of fire conditioned on the independent variables. In abstract terms, the algorithm assumes that all available climate and fire occurrence data can be spatially and temporally co-registered, so as to produce a temporal sequence of data grids for each data source (one grid per day for each data source) over a specific date interval covering several years. In these terms, the independent variables for the statistical model correspond to the data values associated with a particular grid cell location on a specific date, and the independent variable corresponds to the presence or absence of fire within the next N days at that same grid cell location (a binary value).

Important details of the algorithm include a description of the particular variables chosen, the preprocessing steps used to standardize the data, and the particular statistical modeling approach used. The choice of variables was guided in part by a visual analysis of the fire occurrence data, which showed that both land cover type (e.g., grasslands versus coniferous forests) and time of year were important factors in determining fire frequency. For this reason, the algorithm parameters were chosen to produce a unique model for each combination of 17 different land cover types and the 12 different months in the year. The main preprocessing steps involved co-registering data values onto a 8 km-resolution grid covering the coterminous United States, creating separate "training sets" for each statistical model based on land cover type and month of the year, and rescaling values for the independent variables

in each training set to have zero mean and unit variance. More details regarding preprocessing are given in the next section.

The choice of logistic regression as the statistical modeling approach was based upon an extensive study comparing the out-of-sample forecasting accuracy of models based on logistic regression, binary classification trees, and Support Vector Machines (Bonnlander 2005). Results of the comparative study indicate that logistic regression provides consistently superior forecasting accuracy over the other two approaches, and it had the extra benefit of being the fastest of the three approaches by several orders of magnitude. The logistic regression implementation used here came from the glmfit() function provided in the Matlab Statistical Toolbox, Release 13.

Two-phase implementation: The implementation of the wildfire prediction algorithm in the testbed underwent two phases. The goal of the first phase was to build the testbed's hardware and system software environment and to replicate the algorithm originally implemented in the IHMC and compare the algorithm performances in the IHMC machine and the testbed machine. During the first phase, the source code, written in Matlab, and the input data sets was rehosted to the testbed's head node and its local file system. The goal of the second phase was to investigate the performance and feasibility of the algorithm in an operationally-relevant environment with distributed parallel computing and EOS' production data products. During the second phase, most of the data were obtained from RDS through its subscription to operational NASA DAACs. The data were stored in the Xsan administered RAID system in the testbed. The parallelization was achieved through the use of the Matlab Distributed Computing Engine (MDCE). With MDCE, one could start a job manager to manage a number of workers, which perform computation tasks. When multiple workers were started and run at different computing nodes in a computing cluster to collectively perform a subset of a larger computing task, it achieved the goal of parallelizing the computing task. The fire prediction algorithm involved training and building hundreds of models for different land surface types and at different prediction time frames. The Matlab code was written such that model building process could be performed for multiple models or for any subset of models. The code also employed a lock mechanism which would lock a task being processed so that an available computing process could skip a locked task and proceed to the next task. These coding techniques made MDCE very suitable to achieve parallelization. MDCE licenses were installed in four of the six dual-processor computing nodes, which allowed a maximum of eight workers to be run at the same time.

Source Data used in Phase I: Source data used in the algorithm in IHMC and during the first phase of the testbed included four types of information: weather parameters, vegetation condition parameters, national fire occurrence database, and fuel code map.

a) Weather parameters: the weather parameters include four Global Surface Summary of the Day (GSSD) variables from National Climate Data Center

(NCDC). They were daily minimum air temperature (TMIN), maximum air temperature (TMAX), precipitation (PRECIP), and vapor pressure deficit (VPD). There were about 6000 observation points globally, among which about 1200 were in the continental US.

b) Vegetation condition parameters: the vegetation condition parameters included leaf area index (LAI) and Fractional Photosynthetically Active Radiation (FPAR) derived from MODIS measurements. LAI and FPAR were two of the standard MODIS land products. They were available at 1 km spatial resolution.
c) Fire occurrence information: fire occurrence data included the location and size of fire for a specific day. These data were obtained from the National Interagency Fire Management Integrated Database (NIFMID) managed by USFS. The time coverage of the data was from 1986 to 2004, among which 2004 data were 90% complete.
d) Fuel code map: the fuel code map included 24 fuel types among which 20 were vegetation types (Burgan et al. 1998). The map was developed primarily based on the 1-km land cover map derived from AVHRR NDVI by Loveland et al. 1991.

All data were co-registered to gridded 8-km resolution matrices in Lambert Azimuthal equal area projection with center latitude being 45.0 degrees and central meridian being –100.0 degrees. Each grid matrix has 361 rows and 573 columns, covering the entire conterminous US. Both weather and vegetation condition parameters were preprocessed by the Terrestrial Observation and Prediction System (TOPS) project at AMES (http://www.ntsg.umt.edu/tops/webpages/right/flowchart/index.php). MODIS data were aggregated from 1 to 8 km while the weather data were interpolated to 8 km grid using techniques described in Jolly et al. (2005). These data covered time periods between March 5, 2000 and December 31, 2004. The National Interagency Fire Management Integrated Database (NIFMID) provides records of fires occurring on USFS land that required suppressive action for the years 1986–2003 (USDA 1993), including fire location, ignition date and final fire size. The fuel code map, given at 1-km resolution with the same map projection, was aggregated to 8-km resolution using Matlab code. Preprocessing code was written to convert the data into Matlab matrices containing standardized values for independent variables.

Source Data used in Phase II: During the second phase, data obtained directly from the operational NASA DAACs were used to replace some of the input data used in phase I. These included MODIS snow/ice, FPAR/LAI, land cover, Thermal/fire, and TRMM 3-hr gridded precipitation data. Among the above operationally available data, the MODIS FPAR/LAI product was also used in the phase I implementation, but during phase II the original preprocessed data were replaced by data directly from the NASA DAAC. For the other data products, MODIS snow/ice was used as a new prediction variable; MODIS land cover was used to

replace the fuel code; TRMM precipitation was used to replace the original precipitation data; and the MODIS thermal/fire was used to replace the dependent fire occurrence variable. The minimum and maximum air temperature and the vapor pressure deficit data used in phase I were still used in Phase II model training and forecast. It was originally planned to also replace these preprocessed data using NASA operational data such as MODIS land surface temperature and atmospheric profile products. After some initial implementation tests, it was determined that the resources required would exceed those available to the project.

Preprocessing of NASA operational data: The operational NASA data were in different resolutions, both spatial and temporal, and in different coordinate reference systems (CRS) as compared to the data sets used in Phase I. Preprocessing was needed to transform the operational data into a form that could be input into the IHMC algorithm. The preprocessing functionalities included a) CRS transformation to convert MODIS sinusoidal and TRMM geographic CRS to Lambert Azimuthal Equal Area projection; b) spatial resolution transformation to resample/interpolating 500-m (for the snow/ice data) and 1-km MODIS data and the 0.25-degree TRMM data into 8-km grid; c) temporal resolution transformation to convert TRMM 3-h measurements to daily values; d) mosaicking of multiple MODIS tiles into one single coterminous US grid; and d) file reformatting to convert MODIS HDF-EOS and TRMM native HDF formats to generic binary format. Although it was possible to develop one single tool for all the products because the many aspects of the preprocessing were the same for different products, it was decided that one tool for each product was appropriate because it was easier to configure them for individual products. These tools were placed in the training/prediction process flow so that fire prediction models could be retrained when needed and be used to make real-time or near real-time forecast when new data products become available. Preprocessing of all the five operational products added less than 40 seconds overhead to the model forecast.

Construction of wild fire prediction models: After all input data were preprocessed into a co-registered 8-km grid, they were separated into independent and dependent variables. The independent variables for a given day, T, and a given year, Y, were the following:

a) TMIN, TMAX, VPD, PRECIP, and SNOW on day T in year Y,
b) Averages of TMIN, TMAX, VPD, PRECIP, and SNOW over the days [T-7,T-1] in year Y,
c) Averages of TMIN, TMAX, VPD, PRECIP, and SNOW over the days [T-30,T-1] in year Y,
d) FPAR on day T-1 in year Y,
e) LAI on day T-1 in year Y, and
f) Number of fires in the previous year (all of Y-1)

Where TMIN, TMAX, VPD, PRECIP, and SNOW are minimum temperature, maximum temperature, vapor pressure deficit, precipitation, and percent snow cover, respectively.

The corresponding dependent variables were the following:

a) Fire occurrences (zero or one) in the days [T,T+29] for 30-day models
b) Fire occurrences (zero or one)in the days [T,T+6] for 7-day models
c) Fire occurrences (zero or one) in the day T for 1-day models

That is, for an N-day model, if no fire occurred in the days [T, T+N-1], the dependent variable would have a value of zero, and a value of one otherwise.

The independent and dependent variables were used to train three model types, each for 1-, 7-, and 30-day prediction. If required input data on and before day D-1 are available, the models can predict fire potential respectively, for day D, for a 7-day period staring from D (i.e., [D, D+6]), and for a 30-day period staring from day D (i.e., [D, D+29]). The model building, i.e., the generation of the logistic regression models, is done by calling related Matlab function glmfit(), which is provided in Matlab's Statistics Toolbox.

Because the behavior of fires is different at different times of a year and over different land cover types, one single model is not likely to predict fire at all times and locations. The models were separately built for different land cover types and for each of the 12 months in a year. The MODIS land cover product identifies 17 different land cover types, among which 14 occur in the 2001 data used in the tests reported here. Thus, a complete model building process generates 504 models (i.e., 3 model types; 14 land cover types; 12 months; $3 \times 14 \times 12 = 504$).

10.4.2 The Testbed Hardware and Software Configuration

The testbed was built using an Apple G5 Xserve cluster. The head node of the cluster consisted of dual 2.0 GHz G5 processors, 4 GB of DDR SDRAM and two 250 GB hard disks. There were five cluster nodes, each having dual 2.0 or 2.3 GHz G5 processors with 2 GB DDR SDSRAM and 8 GB hard disk. The combined peak performance of all 12 processors was 103 GFLOPS. The storage system consisted of five RAID arrays with total capacity of 22 Terabytes. The RAID arrays and cluster nodes were interconnected via 2 Gbps Fiber Channels. A gigabyte Abilene network connected the testbed, Remote Data Storage (RDS) facility, the NASA DAACs, and other resources. The RDS contained a transient data storage system of 47 TB, of which 2 TB were allocated to the testbed activities, and a persistent data storage system of 185 TB. The subscription to NASA Data Pool for data products needed by the wildfire prediction algorithm and a notification mechanism to the testbed were set up in RDS. The automatic notification processing and data pulling from RDS to the testbed were implemented in the testbed machine. Several standard open interface protocols and proprietary software were used for data transfer between the testbed and RDS machines and NASA data pools, which included the Nirvana Storage Resources Broker (SRB) software, GridFtp, and Open Geospatial Consortium Web Coverage Service. The operating system of the testbed

was Tiger OSX version 10.4 and its RAID system was administrated using the Xsan version 1.1. The software used to implement the wildfire prediction algorithm was MDCE.

10.5 Results

10.5.1 Computation Speed

As mentioned before, prediction models were built for different land surface types and different model types (i.e., time frames). The MODIS land cover product identified fourteen different vegetative cover types. For each prediction type, 168 models, each for a particular month and a particular vegetative surface type, could be built. With three model types providing 1-day, 7-day, and 30-day predictions, a total of 504 were generated in the testbed. Table 10.1 lists times used to generate the three model types, each containing 168 individual models. In the table, the time for the 30-day model included the time used in the variable preparation, which generated independent and dependent variable arrays from input grid data. The time needed to complete this preparation was 7.1 h using a single worker node. If this time had been deducted, the training time for the single worker 30-day model would have been 15.8 h, which was equivalent to the other two model types.

It is to be noted that the times used in 8-worker distributed computing were less than one eighth of the time used in the single-worker computing. Two factors contribute to these results. The first is that the single worker sessions were performed in the head node of the testbed, on which many other processes from other projects were also running. The second is that the three cluster nodes, on which six of the eight workers were running, were 2.3 GHz G5 dual-processor machines while the head node was a 2.0 GHz G5 dual-processor machine.

Once a model was constructed, the time needed to generate forecast result for the entire conterminous US, at 8-km grid cell size, was about 12 s for one model prediction using the head node. This time included reading the input variables from disk files and writing prediction result to the output disk files. With less than 40 s of preprocessing from NASA operational data included, obtaining the three predictions (i.e., 1-day, 7-day, and 30-day) each day would take less than 2 min with a single node.

Table 10.1 Times used in generating the three types of prediction models (in hours)

	Single worker	Eight workers
1-day model	16.5	1.6
7-day model	16.3	1.5
30-day model	22.9	2.8

Note: The time in the 30-day model includes preprocessing time. See text.

The prediction computations scale linearly with the number of grid cells. Thus, if prediction models are used at a 1-km resolution, the time needed to generate the three predictions will be about 2 h since the number of grid cells increase by a factor of 64. This would still meet operational daily forecast requirements. Of course, the time used to train and build the forecast models would be much longer than those shown in Table 10.1.

10.5.2 Forecast Results

The forecast results by the 1-day, 7-day, and 30-day models were probabilities of fire occurrences in the next day, within the next 7 days, and within the next 30 days, respectively. Although the probabilities, floating point values ranging from 0.0 to 1.0, indicated the likelihood of fire occurrences, they could not be thresholded to definitive 1 or 0 values to predict if there would or would not be wildfires. Therefore, it was not possible to directly compare the forecast results to the known fires to determine the accuracy of the forecasts. Hence, visual comparisons between forecast images and known fire images and the Receiving Operating Characteristic Curve (ROC) analyses were used to assess the forecasts. A few examples of predicted images and ROC curves are presented here to show the forecast results. Figures 10.5 and 10.6 are 30-day forecast results for the winter and summer seasons and Fig. 10.7 is a 7-day forecast result for the spring season, all in 2004. The known fire occurrences are plotted as black dots in the figures. The color bar at the bottom of each figure shows the natural log values of the forecast probabilities. These figures

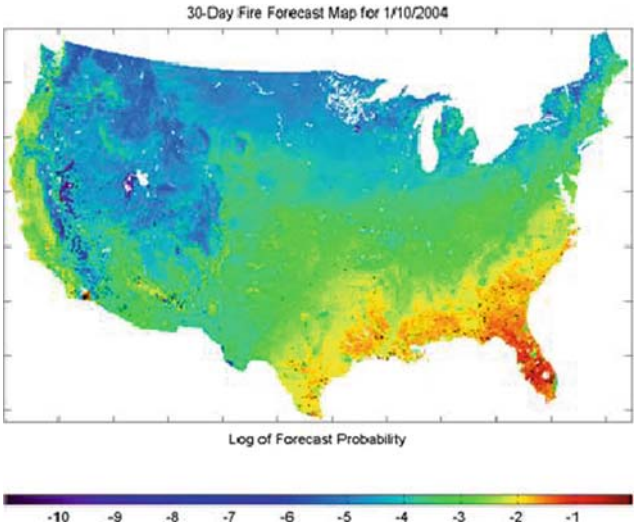

Fig. 10.5 30-day forecast for 1/11/2004–2/9/2004

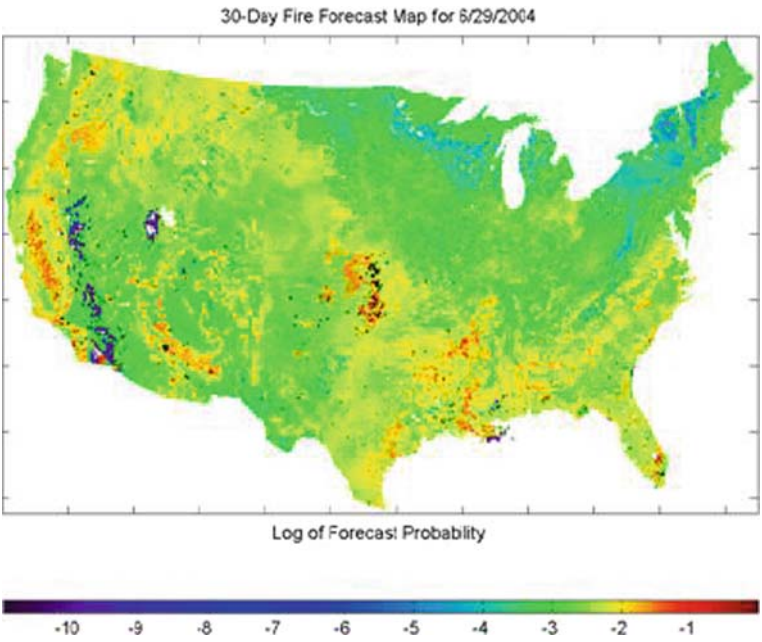

Fig. 10.6 30-day forecast for 6/30/2004–7/29/2004

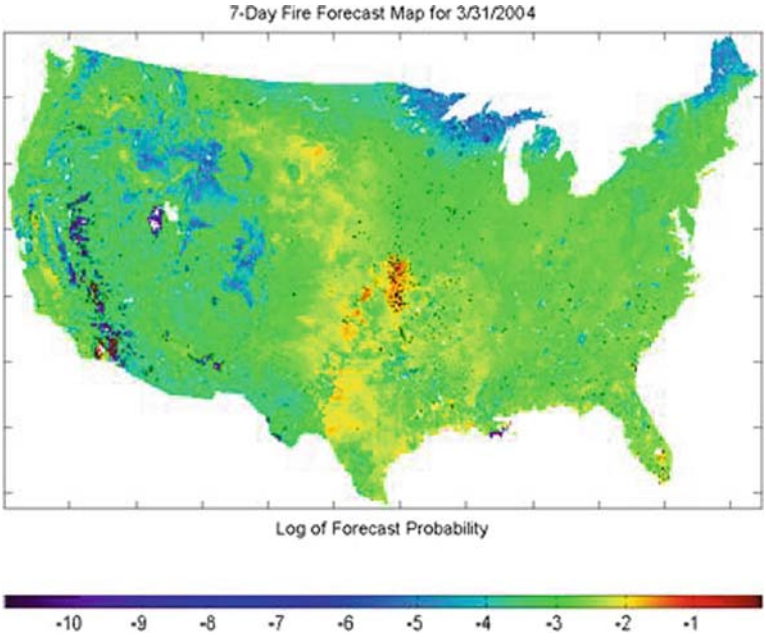

Fig. 10.7 7-day forecast for 4/1/2004–4/7/2004

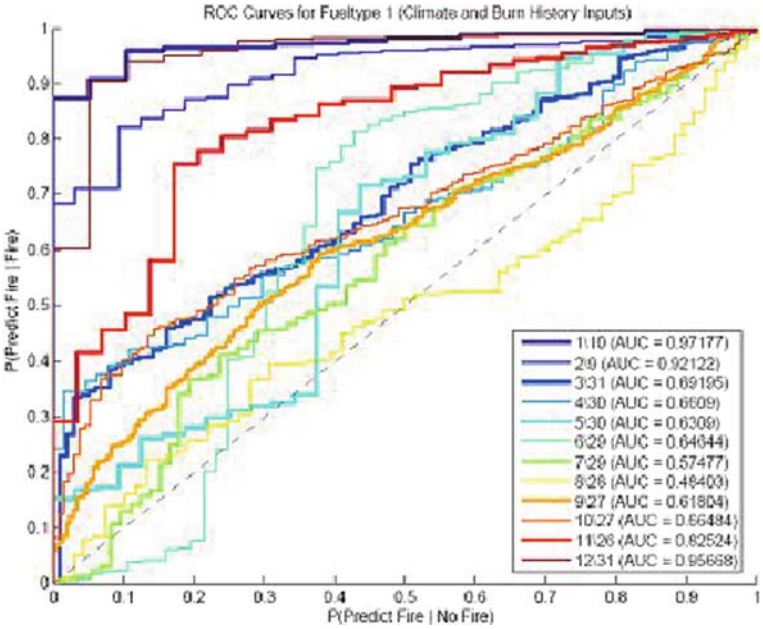

Fig. 10.8 30 day model ROC curves for the evergreen needle leaf forest

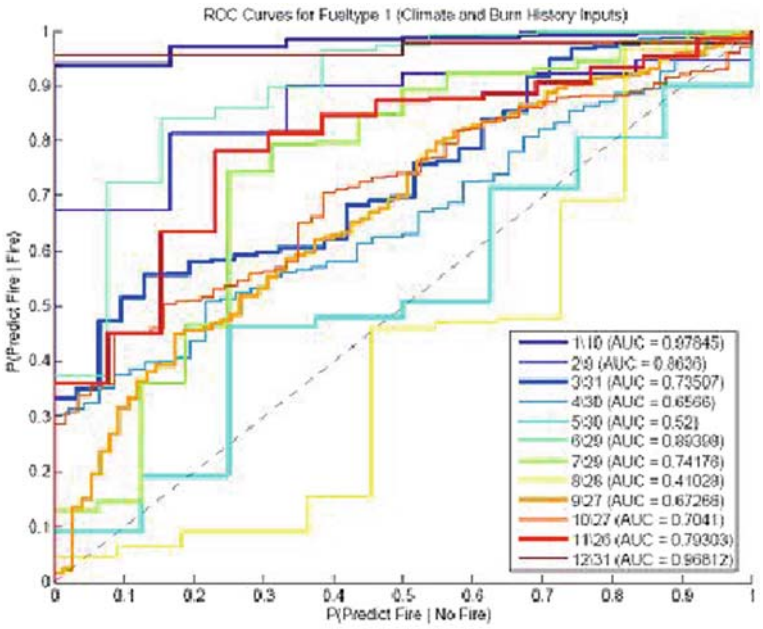

Fig. 10.9 7-day model ROC curves for the evergreen needle leaf forest

indicate that most known fire occurrences, which were not used in model training, fall into the high probability areas. The results demonstrate that the forecasts are visually satisfactory.

More quantitative assessments of the forecasts can be performed by ROC analyses. ROC indicates how the forecast probability separates positive (fire) samples from negative (no-fire) samples. The larger the area under the ROC curve (AUC), the better the probability separates the samples. The maximum AUC is 1.0, indicating perfect separation between positive and negative samples. An AUC value of 0.5 is random prediction while values smaller than 0.5 indicate worse than random prediction. The smaller-than-0.5 case usually occurs when there are not enough samples (Hopley and Van Schalkwyk 2007). Figures 10.8 and 10.9 are ROC graphs, respectively, for the 30-day and 7-day model forecasts calculated from the evergreen needle leaf forest land cover type. These graphs show that the model forecast results are much better than random prediction in most cases. The average AUC values over twelve months for these two graphs are 0.72 and 0.74, respectively.

10.6 Observations and Conclusions

We have discussed the implementation of a testbed for an intelligent archive to demonstrate the feasibility of using data mining algorithms on a large scale with remotely sensed data in an operationally relevant environment. The testbed used for this work consisted of hardware and software in a distributed environment that was distinct from, but interfacing with, NASA's operational Distributed Active Archive Centers, so that there would be no interference with the operational systems. Several of the previously identified key functions of an Intelligent Archive were exercised through this testbed. A data mining algorithm employing logistic regression was used to develop a fire prediction model from time series of a variety of remotely sensed data and derived products. The results from the testbed were encouraging from several points of view. Some of the lessons learned and observations are given below from the points of view of: Science/Algorithm and Execution Efficiency.

10.6.1 Science/Algorithm

Algorithm and parameter selection is science-driven. This implies that substantial, on-going, active guidance by science subject matter experts is essential. Automation can greatly reduce the workload of the trained investigator, but cannot replace the investigator's expertise.

Interpretation of data mining results requires domain expertise. The development and validation team must include members with a broad range of technical and scientific skills appropriate to the problem. A selected mix of algorithm experts, domain scientists, and statisticians need to be involved, perhaps on an ad hoc basis in the development and validation phase.

Logistic Regression makes interpretation of results relatively easy. A significant consideration in the selection of logistic regression (given that predictive performance was not sacrificed) was the heuristic significance of the model produced by this algorithm. The parameters produced by the Logistic Regression model have a natural meaning which can be read, interpreted, and understood by a knowledgeable person without the need for significant additional transformation.

Correlated variables complicate interpretation of results. The logistic regression model takes a weighted sum of its input values to produce an output value, using weights derived from regression over a sample. If the sample has a pair of highly correlated variables, the resulting model will contain a pair of weights whose sum is well determined, but whose individual values are not. Therefore, analysis of the influences of individual variables is limited to those that are uncorrelated, and the remaining variables should be analyzed in groups.

10.6.2 Execution Efficiency

Data mining is feasible on large data volumes. The conceptual architecture for the IA-KBS envisions an ongoing algorithm development and validation process; and the testbed results confirm that this process is computationally manageable. Given scientific collaboration indicated above, the experience from the testbed suggests that the data mining development, validation, and extension of such algorithms is well within the state of the art and the computational resources of commercial off-the-shelf systems.

Mined models can be computationally efficient. Near real-time event detection (or prediction) is well within the current, modest computer system capabilities, based on the timings indicated in Sect. 10.5.1. This also suggests that content-based retrieval is very feasible.

Performance & flexibility of pre-processing is important. A flexible and powerful development environment is important because the preprocessing code can be much larger than the actual data mining code. Efficiency of the preprocessing code is important because this code must run not only in the model training phase (which can be performed against a sample and without significant time constraints), but also in the model execution phase (in near real-time).

Acknowledgment The authors would like to thank the following individuals for their assistance and contributions of ideas to this work: G. McConaughy and C. Lynnes (NASA Goddard Space Flight Center-GSFC) – IA-KBS concepts, S. Morse (SOSACorp) – Intelligent Data Understanding (IDU) research assessment and testbed conceptual design, L. Di (GMU) – IA-KBS concepts, X. Li (GMU) – modeling and forecasting software execution on the testbed, T. Chu (IHMC) – assessment of software porting to testbed, and P. Smith (MacDonald, Dettwiler & Associates, Ltd.), B. Koenig (Electronic Data Systems Corporation - EDS) and C. Yee (Raytheon) – support for RDS/testbed interfaces. They would also like to thank C. Bock, D. Lowe and M. Esfandiari (Earth Science Data and Information System ESDIS Project, NASA GSFC) for their encouragement and support.

The data used in this study were obtained from several sources as indicated in Sect. 10.4.1. The sources include NOAA's National Climate Data Center, US Forest Service, NASA's Ames Research Center, and three of NASA's EOSDIS Data Centers – Land Processes DAAC, National Snow and Ice Data Center and the Goddard Earth Science Data and Information Services Center.

Acronyms

ACCESS	Advancing Collaborative Connections for Earth System Science
AUC	Area under the ROC curve
AVHRR	Advanced Very High-Resolution Radiometer
CRS	Coordinate Reference Systems
DAAC	Distributed Active Archive Center
DCE	Distributed Communication Environment
DDR	Double Data Rate
ECHO	EOS ClearingHOuse
EDC	EROS Data Center
EDS	Electronic Data Systems
EOS	Earth Observing System
EOSDIS	Earth Observing System Data and Information System
FPAR	Fractional Photosynthetically Active Radiation
FTP	File Transfer Protocol
GFLOPS	Giga (10**9) Floating Point Operations Per Second
GMU	George Mason University
GSFC	Goddard Space Flight Center
GSSD	Global Surface Summary of the Day
HDF	Hierarchical Data Format
IA	Intelligent Archive
IA-KBS	Intelligent Archive in the Context of a Knowledge Building System
IDU	Intelligent Data Understanding
IHMC	Institute for Human and Machine Cognition
LAI	Leaf Area Index
LP DAAC	Land Processes DAAC
LSDM	Large-Scale Data Mining
MCAT	Metadata Catalog
MDCE	MatLab Distributed Computing Engine
MEaSUREs	Making Earth Science Data Records for Use in Research Environments
MODIS	Moderate-Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NCDC	National Climate Data Center
NDVI	Normalized Difference Vegetation Index
NOAA	National Oceanic and Atmospheric Administration
PRECIP	precipitation
RAID	Redundant Array of Independent Disks
REASoN	Research, Education and Applications Solutions Network
RDS	Remote Data Storage
ROC	Receiver Operating Characteristic
SDRAM	Synchronous Dynamic Random Access Memory

SRB	Storage Resources Broker
TB	Terabyte
TCP/IP	Transmission Control Protocol/Internet Protocol
TMIN	Temperature, Minimum
TMAX	Temperature, Maximum
TOPS	Terrestrial Observation and Prediction System
US	United States
USFS	United States Forest Service
VPD	Vapor Pressure Deficit

References

Bonnlander BV (2005) "Statistical Forecasts of Wildfire: A Baseline Approach"; (2005) CMU Laboratory for Symbolic Computation Technical Report #172

Burgan R, Klaver R, Klaver J (1998) Fuel models and fire potential from satellite and surface observation. International Journal of Wildland Fire 8:159–170

Clausen M, Lynnes C (July 2003) Virtual Data Products in an Intelligent Archive, White Paper prepared for the Intelligent Data Understanding program, http://daac.gsfc.nasa.gov/intelligent_archive/presentations.shtml

Hopley L, Van Schalkwyk J (March 3, 2007) "The magnificent ROC (Receiver Operating Characteristic curve)", http://anaesthetist.com/mnm/stats/roc/Findex.htm

Isaac D, Lynnes C (January 2003) Automated Data Quality Assessment in the Intelligent Archive, White Paper prepared for the Intelligent Data Understanding program, http://disc.gsfc.nasa.gov/IDA/

Isaac D, McConaughy G (September 2004) "Intelligent Archives in the Context of Knowledge Building Systems: Data Volume Considerations", White Paper prepared for the Intelligent Data Understanding program, http://disc.gsfc.nasa.gov/IDA/

Jolly W, Graham J, Michaelis A, Nemani R, Running S (2005) A flexible, integrated system for generating meteorological surfaces derived from point sources across multiple geographic scales. Environmental Modeling & Software, 20:873–882

Loveland T, Merchant J, Ohlen D, Brown J (1991) Development of a land-cover characteristics database for the conterminous US. Photogrammetric Engineering and Remote Sensing, 57:1453–1463

McConaughy G, McDonald K (September 2003) "Moving from Data and Information Systems to Knowledge Building Systems: Issues of Scale and Other Research Challenges," White Paper prepared for the Intelligent Data Understanding program, http://disc.gsfc.nasa.gov/IDA/

Morse S, Isaac D, Lynnes C (January 2003) "Optimizing Performance in Intelligent Archives," White Paper prepared for the Intelligent Data Understanding program, http://disc.gsfc.nasa.gov/IDA/

Morse S, Yang W (October 2004) "A Conceptual Specimen Architecture for an Intelligent Archive in a Knowledge Building System," White Paper prepared for the Intelligent Data Understanding program, http://disc.gsfc.nasa.gov/IDA/

NASA (February 3, 2005) "Evolution of EOSDIS Elements," Study Team Briefing to NASA, http://eosdis-evolution.gsfc.nasa.gov/

Ramapriyan H, McConaughy G, Morse S, Isaac D (August 2004) "Intelligent Systems Technologies to Assist in Utilization of Earth Observation Data," presented at Earth Observing Systems IX, SPIE Meeting, http://disc.gsfc.nasa.gov/IDA/

Ramapriyan H, Isaac D, Yang W, Morse S, (2005) "Large Scale Data Mining to Improve Usability of Data – an Intelligent Archive Testbed," IGARSS, Seoul, Korea

Ramapriyan H, Behnke J, Sofinowski E, Lowe D, Esfandiari M (2009) "Evolution of the Earth Observing System (EOS) Data and Information System (EOSDIS)," Chapter 7, Standards Based Data and Information Systems for Earth Observations (Eds) Di L and Ramapriyan H, Springer-Verlag, New York.

USDA Forest Service (1993) National Interagency Fire Management Integrated Database (NIFMID) reference manual; US Department of Agriculture, Forest Service, Fire and Aviation Management. Washington, DC, USA, p 43

Chapter 11
Semantic Augmentations to an ebRIM Profile of Catalogue Service for the Web

Peng Yue, Liping Di, Peisheng Zhao, Wenli Yang, Genong Yu, and Yaxing Wei

11.1 Introduction

A geospatial catalogue service provides the capabilities to advertise and discover shared data and services over the Web. Description Information (Metadata) for data and services is stored and organized in a catalogue service to allow search. The Open Geospatial Consortium (OGC)'s Catalogue Service for the Web (CSW) has been defined by industry consensus. It is a standard interface to online catalogs of geographic information and Web-accessible geoprocessing services. CSW specifies interfaces, HTTP protocol bindings, and an abstract information model for defining application profiles required to publish and access digital catalogues of metadata for geographic data, services, and related resource information (Nebert and Whiteside, 2005). OGC has developed and recommended an ebRIM profile of CSW for implementing a catalogue service (Martell, 2004).

While CSW greatly facilitates the discovery of data and services, the current discovery process is based on matching static keywords, without fully exploring the underlying semantics such as hierarchical relationships among metadata entities. Semantic augmentations to CSW, can improve the ability to discover data and services. Ontology has been commonly used to represent semantics in computer science. An ontology is a formal, explicit specification of a shared conceptualization that provides a common vocabulary for an area and defines the meaning of terms and the relations between them (Gruber, 1993). By formally conceptualizing metadata for data and services in ontologies, the semantics of metadata can be explicitly defined. Web Ontology Language (OWL) (Dean and Schreiber, 2004), the standard Web ontology language recommended by W3C, provides the ability for explicit semantic representation. OWL Service Ontology (OWL-S) (Martin et al., 2004), an OWL based ontology for Web services, supports the description of service capabilities. These explicit specifications make the semantics of geospatial

P. Yue (✉)
State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; Center for Spatial Information Science and Systems (CSISS), George Mason University, Greenbelt, MD 20770, USA
e-mail: geopyue@gmail.com

L. Di, H.K. Ramapriyan (eds.), *Standard-Based Data and Information Systems for Earth Observation*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-540-88264-0_11,

data and services machine-understandable, allowing use of semantics for flexible discovery of geospatial data and services. This chapter explores the semantic representation of geospatial data and services to use the semantic relationship defined in OWL/OWL-S for semantic search in CSW.

To use OWL/OWL-S in the CSW discovery process, the semantic relationships in OWL/OWL-S must to be stored in CSW. Two industry models exist for information registry: Universal Discovery Description and Integration (UDDI) (OASIS, 2004) and Electronic Business Registry Information Model (ebRIM) (OASIS, 2005). UDDI deals only with services: its registry model is not flexible enough for data registration compared to ebRIM. The ebRIM profile for CSW implementation introduces an ebRIM-based catalogue information model for publishing and discovering geospatial information. The ebRIM model is a widely adapted information model that defines types of objects stored in a registry as well as the relationships among these object types. Geospatial data and services can be registered using this model, following the geospatial metadata standards. However, the current CSW specification does not take into account the registration of the semantic information in a CSW. Because ebRIM is a general and extensible registry information model that can be extended through its class, slot and association elements, these elements are extended to allow registration of the semantic information for geospatial data and services.

This work has implemented a semantically-augmented searching component, serving as a middleware between the CSW and requestors. It performs semantic match to improve the recall and precision of data and services discovery. Chaining and execution of geospatial Web services provides a flexible yet powerful way to derive high-level geospatial information from lower-level inputs through real-time integration of interoperable geospatial services and data. The work incorporates an automatic, "DataType"-driven service chaining process in the semantically-augmented searching component. The chaining process produces an executable composite service to generate an on-demand geospatial product corresponding to user's requirements when there is no semantically-matched data available.

The chapter is organized as follows: Section 11.2 briefly introduces the ebRIM-based information model in CSW and an implementation by GMU LAITS. Section 11.3 gives the related work on semantics-enabled service registries. In Sect. 11.4, the semantic description for geospatial data and services is provided. Section 11.5 describes how semantics defined in OWL/OWL-S ontologies can be represented in CSW. Section 11.6 describes how semantically augmented search functions are implemented. Finally, Sect. 11.7 concludes the chapter.

11.2 The ebRIM-based Information Model in CSW

11.2.1 CSW Services Specification

According to the CSW specification, any implementation of CSW must consist of two components.

(1) A catalogue abstract information model, which consists of catalogue query language and core catalogue schema.

Catalogue query language provides a minimal abstract query language, which must be supported by all compliant OGC Catalogue Services. It supports nested Boolean queries, text matching operations, temporal data types, and geospatial operators. The minimal query language syntax is based on the SQL WHERE clause in the SQL SELECT statement. Implementations of query languages that are transformable to the OGC_Common Catalogue Query Language include the OGC Filter Specification (Vretanos, 2005).

The core catalogue schema provides the basic metadata information model. It consists of core queryable properties and core returnable properties. Different geospatial applications or CSW profile implementations might adopt different metadata information models to describe the metadata. Each type of metadata information model must support the core queryable properties and core returnable properties. The support for core queryable properties enables those clients compatible with the OGC CSW specification to access different implementations of OGC CSW services and implement simple cross-profile discovery. At the same time, the support for core returnable properties provides a common representation for the geospatial information advertised in different implementations of OGC CSW services, thus facilitate information exchange among different geospatial applications and user groups.

(2) HTTP protocol binding. CSW supports access via HTTP protocol. Clients and servers interact using a standard HTTP protocol request-response model. That is, a client sends a request to a server using HTTP, and expects to receive a response to the request or an exception message. Seven operations are defined in the CSW interface: GetCapabilities, DescribeRecord, GetDomain, GetRecords, GetRecordById, Harvest, and Transaction. GetCapabilities, DescribeRecord, GetRecords, and GetRecordById are mandatory for CSW implementation, while the other three are optional. These operations can be divided into three classes:

The first class is service operations, which are operations that a client may use to get the capabilities of a service. GetCapabilities is the only one. It allows CSW clients to retrieve service metadata from a server.

The second class is discovery operations, which a client may use to determine the information model of the catalogue and query catalogue records. This class consists of DescribeRecord, GetDomain, GetRecords, and GetRecordById. DescribeRecord allows a client to get the description of elements in the information model. GetDomain is used to get the range of values of a metadata record element or request parameter. The GetRecords operation is the main operation for resource discovery. Resource discovery in the general model can be characterized as two operations: *search* and *present*. They are combined in the GetRecords operation of the HTTP protocol binding. GetRecordById operation is an implementation of the *present* operation. It retrieves the metadata records using their identifier. In the work here, the GetRecords operation implements search functionalities.

The third class is management operations, which are used to create or update records in the catalogue. The Harvest and Transaction operations are in this class.

The Transaction operation defines an interface for creating, modifying, and deleting catalogue records. It can be treated as a "push" operation. Harvest is a "pull" operation. It refers only to the data to be inserted or updated in the catalogue. Catalogue services need to resolve the reference, fetch that data, and process it into the catalogue. We use the Transaction operation in our implementation to register semantics.

11.2.2 An ebRIM Profile of CSW

The CSW specification provides a framework for implementing application profiles. OGC has developed and recommended an ebRIM profile of Catalogue Service for the Web for implementing a catalogue service. The UML style graph of Fig. 11.1 shows a high-level view of the ebRIM-based catalogue information model.

The core class in the ebRIM model is the "RegistryObject". This is an abstract base class used by most classes in the model. It provides minimal metadata for registry objects. It also provides methods for accessing related objects that provide additional metadata for the registry object. "Slot" instances provide a flexible way to add arbitrary attributes to "RegistryObject" instances. "Association" instances are RegistryObject instances that are used to define many-to-many associations between

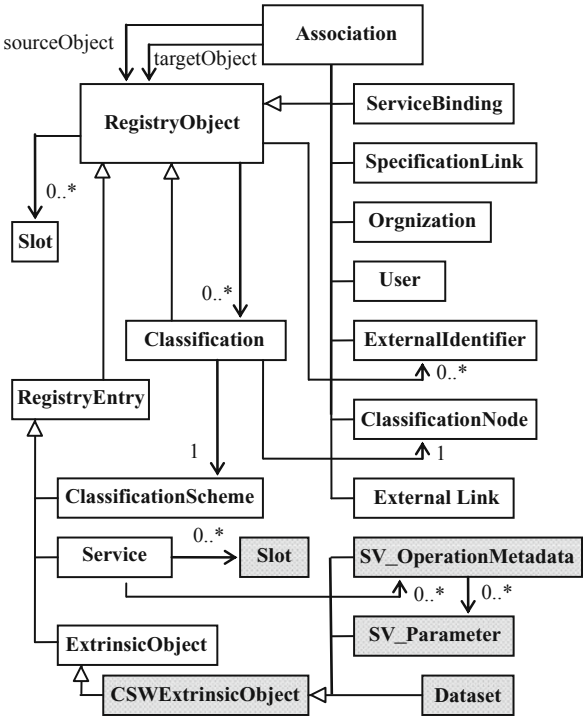

Fig. 11.1 High-level view of catalogue information model

objects in the information model. An Association instance represents an association between a source RegistryObject and a target RegistryObject. Each Association must have an "associationType" attribute that identifies the type of that association.

"ClassificationScheme" instances are instances of RegistryEntry that describe a structured way to classify or categorize RegistryObject instances. The structure of the classification scheme may be defined internally to or externally of the registry, resulting in a distinction between internal and external classification schemes. "ClassificationNode" instances are instances of RegistryObject that are used to define tree structures under an internal ClassificationScheme, where each node in the tree is a ClassificationNode and the root is the ClassificationScheme. Classification trees constructed with ClassificationNodes are used to define the structure of Classification schemes or ontologies.

Instances of "Classification" are instances of RegistryObject that can be used to classify other instances of RegistryObject. Using instances of Classification, instances of RegistryObject can be classified by specific taxonomy values in different schemes (i.e. instances of ClassificationScheme). Thus, a Classification is somewhat like a specialized form of an Association. Classifications could be internal or external depending on whether the classification scheme used is stored internal to the registry or it is external. An internal classification always refers to the ClassificationNode directly by its ID while an external classification refers to the node indirectly by specifying a representation of node value that is unique within the external classification scheme.

The ebRIM model provides a general and standard model for metadata registration information. However, some extension elements are needed, to meet common requirements in the geospatial domain. Using the guidelines of the ebRIM profile for CSW, the GMU LAITS CSW (Wei et al., 2005b) implementation extended ebRIM using international geographic standards: ISO 19115 (ISO/TC211, 2003) (including draft part 2: Extensions for imagery and gridded data) and ISO 19119 (ISO/TC211, 2005). The dark rectangles in the Fig. 11.1 show the extension to the ebRIM.

The ebRIM is extended with ISO 19115 and ISO 19119 in two ways. The first is importing new classes into the ebRIM class tree, deriving new metadata classes from existing ebRIM classes. "ExtrinsicObject"s provide metadata that describes submitted content whose type is not intrinsically known to the registry and therefore must be described by means of additional attributes. A class "CSWExtrinsicObject" is derived from the class ExtrinsicObject to represent all the metadata objects describing objects that may not be intrinsic to the catalogue. The Dataset class is derived from CSWExtrinsicObject to describe geographic datasets. Many new attributes are added to the Dataset class based on ISO 19115 and its draft part 2.

The second way to extend ebRIM is to use Slots to extend an existing class. Every class extended from the RegistryObject class can add Slots into itself. The Service class included in the ebRIM can be used to describe geographic service but the available attributes in the class Service are not sufficient to describe geospatial Web services. Thus, new attributes derived from ISO 19119 are added to the Service class through Slots.

The CSW Service implemented by GMU LAITS follows the OGC CSW specification to ensure generality and interoperability. The query language is implemented using the OGC Filter specification. It supports comparison operators (PropertyIsEqualTo, PropertyIsNotEqualTo, PropertyIsLessThan, PropertyIsGreaterThan, PropertyIsLessThanOrEqualTo, PropertyIsGreaterThanOrEqualTo, PropertyIsLike, PropertyIsNull, and PropertyIsBetween) and spatial operators (BBOX, Within, Touches, Overlaps, and Contains). A client can query the core queryable properties of the CSW. In the internal implementation, the core queryable properties are mapped to the corresponding metadata properties in ebRIM. The response metadata records can be organized according to the users' request, either by following the extended ebRIM model or by following the basic metadata records defined in the CSW specification, where the basic metadata records include the core returnable properties.

11.3 Semantics-Enabled Service Registry – Current Solutions

The service registry plays an important role in helping requestors to find the right services. Web services are catalogued in a registry/broker with their properties and capabilities. As was mentioned in the introduction section, there are two prominent general models for registry services: ebRIM and UDDI. Currently, the search functionality for both of them is limited to the direct match of keywords from metadata; it does not fully use the semantic information, such as hierarchical relationships among metadata entities implicitly embedded in the metadata. Also, no search mechanism is based on the operation/functionality, input/output, and pre/post-conditions of services. Efforts to add semantic information into UDDI or ebRIM to enable a semantics-enabled search are in progress.

11.3.1 Adding Semantics into UDDI

There are efforts in the business world to add semantics to UDDI. Paolucci et al. (2002a) introduce a mapping from OWL-S to UDDI data model. UDDI describes three types of entities:

a. Business Entity, which records contact and owner information;
b. Business Service, which describes one or more specific services that a business provides;
c. Binding Template, which specifies the access point of service.

In addition to these entities, UDDI provides a data structure called TModel that can specify the additional attributes of entities, thus allowing description of the specified ontological concepts. Each service can have one or more TModels that help describe its characteristics. Thus, service capabilities such as function or service input/output can be recorded in the corresponding TModels. Most other efforts use similar mappings; they differ only in the implementation of semantic search.

Three options for semantic search are available.

Option 1: Functionality is created outside of UDDI without any change to the UDDI interface (Paolucci et al., 2002a). The OWL-S Matching Engine is developed to handle the semantics-enabled search. Registration of service semantics consists of the following steps:

a. Advertising services in the form of OWL-S;
b. Using the mapping of the OWL-S profile to the UDDI data model, constructing the UDDI service description using information in the OWL-S and registering it into the UDDI.
c. Getting the reference ID for the service from the result of registration with UDDI, combining it with the capabilities description in the service advertisement (i.e. OWL-S), and storing them into the AdvertisementDB (Advertisement Data Base) component of the OWL-S Matching Engine.

Semantics-enabled service discovery has the following steps:

a. Constructing the service request in OWL-S form.
b. Selection by the OWL-S Matching Engine of the advertisements from the AdvertisementDB, using the output-first and input-second semantic matches (Paolucci et al., 2002b) to compute the level of match to the request.
c. Getting the UDDI records using the reference ID of the matching result and combining them with the advertisement from the matching result as the response.

Because the number of advertisements may be large, the matching process can be extremely time-consuming. The degree of match is pre-computed in the publishing phase of the service. Each ontological concept is indexed with the related services and their match level at input or output. Since the matching information has been pre-computed at the publishing phase, the query phase is reduced to simple lookups in the hierarchical data structure (Srinivasan et al., 2004).

Instead of mapping OWL-S to UDDI structures, Sivashanmugam et al. (2003) introduce the mapping from WSDL-S to UDDI structures, while the design of TModels in the UDDI is still similar to the method above. They enhance the matching ability of Paolucci et al. (2002b) technique by considering the functionality of service operations. First, services are selected using the ontological concepts of functionality, then they are pruned using the input and output match.

Option 2: The functionality is embedded into UDDI with some changes to the UDDI interface to support the semantically augmented query (Akkiraju et al., 2003). The UDDI API schema is extended with a property (RDF:Property) referring to the ontological concepts. The service publishing steps are similar to the Paolucci et al. (2002a) method, except that no AdvertisementDB is maintained. The service discovery steps are as follows:

a. Constructing the service request following the UDDI API schema (at this time the request contains semantic representation according to the schema extension).
b. Getting the set of services filtered according to the standard UDDI schema (the standard UDDI find method can be used).

c. Sending the filtered set of services to the semantic matching engine, to enable a semantic match with the requested ontological concepts. The degree of match is based on the input and output matches. If no match is available, the semantic matching engine will compose services to meet the original request.

Option 3: The functionality is wrapped as an individual external matching service registered into UDDI. In this option, UDDI relays the matching task to the external matching service to enable the different types of matching such as OWL-S, WSDL and UML (Colgrave et al., 2004). The registered service information includes the identification of its appropriate external matching information. The service discovery process consists of three stages:

a. Detecting the need for external matching from the request and taking it as a filter to retrieve the relevant external matching description of services.
b. Looking for available and compatible external matching services and invoking the appropriate external matching service by passing the requirements as well as the filtered service descriptions.
c. Finding the services according to the matched external descriptions.

11.3.2 Adding semantics into ebRIM

In recent studies, OWL elements have been mapped to ebRIM elements (Dogac et al., 2004; Wei et al., 2005a). The basic idea is to use the Classification, Slot, and Association elements in ebRIM to record corresponding OWL classes, properties and related axioms such as subclassOf. However, few studies of registering OWL-S into ebRIM are available. Although OWL-S is mentioned by Dogac et al. (2004), only hierarchical OWL classes inheriting from the OWL-S "Service" class have been explored for registration in ebRIM. The semantics of service instances such as input and output cannot be used in its queries search. In addition, there is no semantic matching engine to enable semantic search, since current OWL reasoners to be used in the matching engine are based on the syntax of OWL and cannot directly run on the ebRIM. Dogac et al. (2005) provide an approach to processing the registered semantics using stored procedures. Using the semantic meaning of ebRIM element constructs corresponding to the OWL constructs, the stored procedures processes the ebRIM construct to achieve a limited reasoning capability. The disadvantage of this approach is that it loses the power of the OWL reasoner.

11.4 Semantics for Data and Services

Metadata for geospatial data includes much descriptive information such as identification, constraints, data quality, spatial/temporal representation, and content (ISO/TC211, 2003). Constructing a semantic representation of the descriptive information is a major research topic. The research described here focuses on the

semantic representation of data content, which involves the conceptualization using OWL for a taxonomy of scientific theme keywords. We refer to such OWL as "DataType" ontology. An example of such ontology derived from the conceptualization of the Global Change Master Directory (GCMD) science keywords is shown in Fig. 11.2. The figure was captured using Protégé (http://protege.stanford.edu/), a freely available tool that can support OWL.

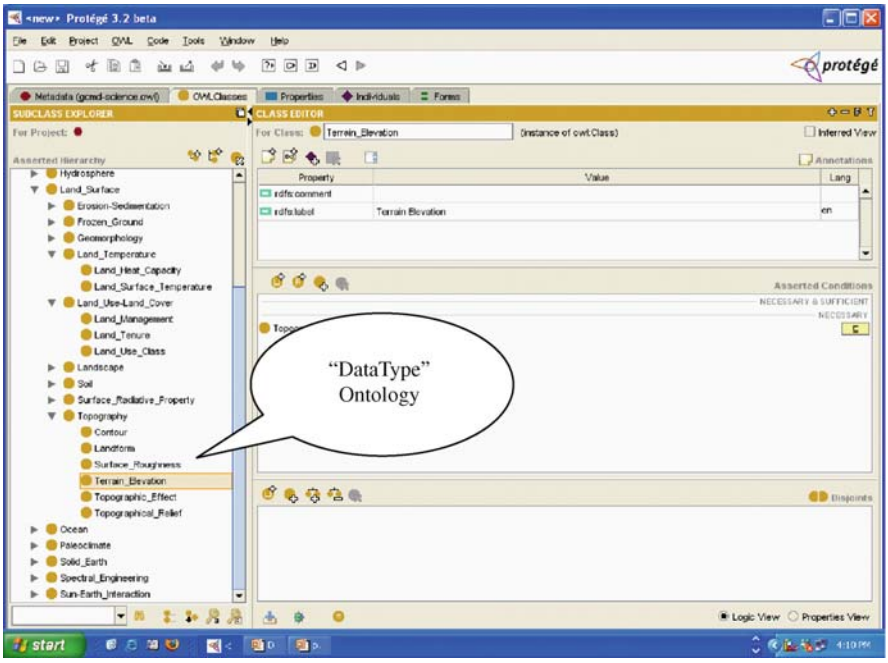

Fig. 11.2 Protégé snapshot GCMD science keywords in OWL

According to ISO 19119, there are three types of service metadata entities (ISO/TC211, 2005):

(1) Service Instance: the service itself, hosted on a specific set of hardware and accessible over a network;
(2) Service Metadata: description of service operation, address, and owner etc.;
(3) Service Type: describes the specific service type, e.g. Web Coverage Service (WCS), Web Coordinate Transformation Service (WCTS) of a service instance.

GCMD provides a comprehensive hierarchical keyword list for services. This keyword list can be conceptualized into "ServiceType" ontology (Fig. 11.3). The "ServiceType" ontology can be used to address the type of service.

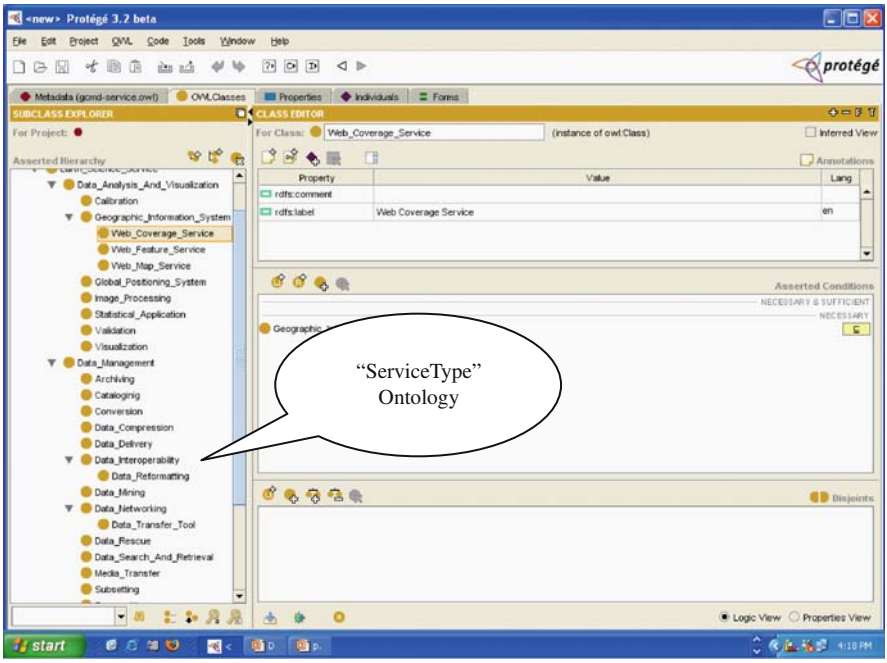

Fig. 11.3 Protégé snapshot GCMD service keywords in OWL

OWL-S as a service ontology is characterized by three modules: Service Profile, Service Model, and Service Grounding. Service Model describes how a service works. Service Grounding shows how to access a service, and Service Profile provides the capabilities of Web services such as services' operation/functionality, input/output, and pre/post-conditions. Thus, the service profile can be used as the basis for service selection. For example, a slope service has the Digital Elevation Model (DEM) input type, and it may have a pre-condition on data format (e.g., HDF-EOS format). On output, it generates a specific data type, the terrain slope type. The OWL-S support for automatic service discovery and composition is useful for real-time production of requested data which otherwise do not exist. For example, if slope data requested by a client does not exist, the service chaining capability may dynamically compose a chain by searching in the semantic CSW for a slope service and the input data needed to feed into it. If the slope service is general enough to be applied to a broad spatial and temporal extent, the composed service chain can be viewed as a "Virtual data product", which can be catalogued in the CSW and be instantiated on demand. OWL-S is still under development, thus the objective is not to incorporate all the metadata for service in OWL-S. Rather, OWL-S is introduced as an augmentation for service description, to enable the production of a "Virtual Data Product" through automatic service chaining. The "DataType" and "ServiceType" ontologies are used in the OWL-S for descriptions such as input/output "DataType", and service classification.

A service instance can be either tightly coupled with a dataset instance, or unassociated with specific data instances, i.e. loosely coupled (ISO/TC211, 2005). Loosely coupled services may have an association with data types instead of specific data instances. They can be described through a Service Profile in OWL-S that advertises a certain type of services with specific input/output data types. In addition to the "DataType" and "ServiceType" ontologies, the "Association" ontologies are included to describe the relationships between services and data. The introduction of association ontology can significantly speed up the reasoning process because it reduces the searching space through the association relationship expressed in the ontology. For example in Fig. 11.4, the "DataType" Terrain_Slope[1] is associated with the "ServiceType" Image_Processing. The searching process for services can then start searching within those services of the service type "Image_Processing", which usually results in quicker searching for the needed service. It serves as an optional search optimization strategy in the service discovery process for "Virtual Data Product". As "DataType" and "ServiceType" ontologies act like a conceptual schema for semantic markups of dataset and service instances, they are called "Geospatial Semantics Schema".

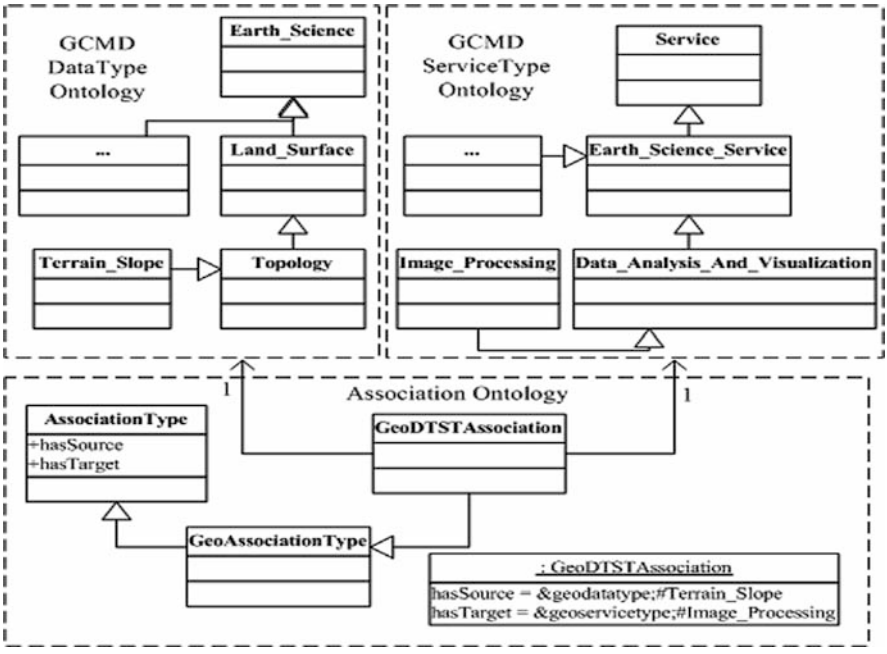

Fig. 11.4 "DataType", "ServiceType", and "Association" ontologies (revised from Yue et al., 2007)

[1]Terrain_Slope is not conceptualized from the GCMD keyword. It is defined by extending entity classes in the GCMD ontology.

11.5 Semantics Registration in CSW

To combine OWL/OWL-S and CSW, OWL/OWL-S must be embedded into the ebRIM-based catalogue information model. Current solutions are provided in Sect. 11.3.2. This study focuses on the application and extension of ebRIM to the geospatial domain. It shows how to register the OWL/OWL-S for geospatial data and services in the ebRIM-based catalogue information model to support semantic search.

The data semantics of each dataset can be annotated in an extended ebRIM element, as described in Sect. 11.4, with the ontology entity class from OWL appropriate to the data thematic information. For service semantics, the most straightforward way is to store the URI from OWL-S in an extended ebRIM element of a service registry object. However, the capabilities advertised in OWL-S must be extracted from the URI by CSW and compared with the search conditions for services, which results in inconvenience. A more general approach is to build a mapping from the service profile of OWL-S to the underlying registry information model, just as the research in the UDDI field did (Sect. 11.3.1). Table 11.1 shows current semantics registration in CSW. Related explanations are listed as follows.

Table 11.1 Semantics registration in CSW

Geospatial Semantics	CSW
Semantics Schema	
"DataType" OWL	ClassificationScheme
"ServiceType" OWL	ClassificationScheme
"Association" OWL	Association
Service Semantics(OWL-S)	
"ServiceType"	ClassificationNode
Input "DataType"	Association
Output "DataType"	Association
Pre-Condition	Service Slot
Post-Condition	Service Slot
OWL-S URI	Service Slot
Data Semantics	
"DataType"	ClassificationNode

(1) The "DataType" and "ServiceType" OWLs are registered in CSW as ClassificationSchemes. A ClassificationScheme can be internal or external. An internal ClassificationScheme allows the registry to validate the subsequent submissions of ClassificationNode and Classification instances in order to maintain the consistency of ClassificationScheme. In an external ClassificationScheme, the structure and values of the taxonomy elements are unknown to the registry. An external ClassificationScheme is suitable for unstable taxonomy, because the update of taxonomy structure and values will not cause much change on the registry.

```
    <ClassificationScheme ... id="urn:uuid:192D785A-4B9D-4be5-B2E9-
01F7FAD21033" isInternal="1" nodeType="UniqueCode">
    <Name> <LocalizedString xml:lang="en-US" charset="UTF-8"
value="GCMD_OWL_DATATYPE"/> </Name>
    <Description><LocalizedString xml:lang="en-US" charset="UTF-8"
value="http://geobrain.laits.gmu.edu/ontology/2004/11/gcmd-
science.owl"/>
    </Description>
 <ClassificationNode id="urn:uuid:09552593-9676-48be-A71E-
6B57C6A30BF3" home="http://laits.gmu.edu:8099/csw" ob-
jectType="urn:uuid:8736A91C-B99B-452D-B638-1CA1CC5A8EC4"
status="Submitted" parent="urn:uuid:192D785A-4B9D-4be5-B2E9-
01F7FAD21033" code="Earth_Science"
path="/GCMD_OWL_DATATYPE/Earth_Science">
 <Name><LocalizedString xml:lang="en-US" charset="UTF-8"
value="Earth_Science"/>
 </Name>
 <Description><LocalizedString xml:lang="en-US" charset="UTF-8"
value="Earth_Science"/>
    </Description>
 <ClassificationNode>
 ...
 </ClassificationNode>
 </ClassificationNode>
    </ClassificationScheme>
```

Fig. 11.5 Internal classificationscheme

Figures 11.5 and 11.6 provide the definitions of internal ClassificationScheme and external ClassificationScheme respectively in XML. The "isInternal" property of internal ClassificationScheme is set as "1", while for external ClassificationScheme, it is set as "0". In the tree-like structure of the internal ClassificationScheme, the root node is the ClassificationScheme instance; the other nodes are the instances of ClassificationNode. Since the structure of an external ClassificationScheme is unknown to the registry, assigning the URI of the OWL in the properties of an external ClassificationScheme is sufficient. Since the GCMD OWL is based on a relatively stable keyword set, the examples described here use an internal ClassificationScheme.

```
    <ClassificationScheme ... id="urn:uuid:192D785A-4B9D-4be5-B2E9-
01F7FAD21033" isInternal="0" nodeType="UniqueCode">
    <Name><LocalizedString xml:lang="en-US" charset="UTF-8"
value="GCMD_OWL_DATATYPE"/> </Name>
    <Description><LocalizedString xml:lang="en-US" charset="UTF-8"
value="http://geobrain.laits.gmu.edu/ontology/2004/11/gcmd-
science.owl"/>
    </Description>
    </ClassificationScheme>
```

Fig. 11.6 External classificationscheme

(2) A new association type "assocGeoDTST" is defined with its sourceObject being a ClassificationNode from the "DataType" ClassificationScheme and its targetObject being a ClassificationNode from the "ServiceType" ClassificationScheme. Each association instance in the "Association" OWL is registered as an Association object under this association type.
(3) Each Service, as an ebRIM RegistryObject, is classified according to the "ServiceType" ClassificationScheme, using the associated ClassificationNode to specify its "ServiceType".
(4) Two new association type "assocServiceOutputDT" and "assocServiceInputDT" are defined, with the sourceObject being a Service object and the targetObject being a ClassificationNode from the "DataType" ClassificationScheme. Each service instance is associated with its input and output "DataType"s through the Association objects under these two new association types. Preconditions and post-conditions can be stored in the extended service slots. Figure 11.7 shows the definitions of "assocServiceOutputDT" and "assocServiceInputDT". The value of objectType property in the ClassificationNode is the Universally Unique Identifier (UUID) of the AssociationType ClassificationScheme, which means that this ClassificationNode is an instance of AssociationType.

In Fig. 11.8, a service instance is defined for a slope service and two association instances are given to address the input and output "DataType" of the slope service. The first association instance shows that the input "DataType" of the slope service is "Terrain_Elevation", using the ID of "Terrain_Elevation" ClassificationNode in the "DataType" ClassificationScheme, while the second association instance shows that the output "DataType" of the slope service is "Terrain_Slope", using the ID of "Terrain_Elevation" ClassificationNode in the "DataType" ClassificationScheme. This classification shows that this service is classified as a "Slope" service type under the "ServiceType" ClassificationScheme.

```
<ClassificationNode ... id="urn:uuid:A80BE843-EB5E-43b1-BBF6-
90B90B8FA973" objectType="urn:uuid:247edbdb-31e8-40bc-97bd-
fd60497deabb" code="assocServiceInputDataType" >
<Name> <LocalizedString xml:lang="en-US" charset="UTF-8"
value="assocServiceInputDataType"/> </Name>
<Description><LocalizedString xml:lang="en-US" charset="UTF-8"
value="association type definition from Service to its input
DataType"/></Description>
</ClassificationNode>

<ClassificationNode ... id="urn:uuid:4EC1D47C-7D70-4b05-A651-
7AA8115FF831" objectType="urn:uuid:247edbdb-31e8-40bc-97bd-
fd60497deabb" code="assocServiceOutputDataType" >
<Name><LocalizedString xml:lang="en-US" charset="UTF-8"
value="assocServiceOutputDataType"/></Name>
<Description><LocalizedString xml:lang="en-US" charset="UTF-8"
value="association type definition from Service to its output
DataType"/></Description>
</ClassificationNode>
```

Fig. 11.7 AssociationType definitions

```
<Service ... id="urn:uuid:19353DEB-CC81-44bb-99D9-
8A1F26FE5F41" >
<Name><LocalizedString xml:lang="en-US" charset="UTF-8"
value="LAITS SLOPE"/>
</Name>
<Description><LocalizedString xml:lang="en-US" charset="UTF-8"
value="Slope calculation Service provided by LAITS at
GMU"/></Description>
<Slot name="OWLSURI" slotType="Service"><ValueList>
<Value>http:// www.laits.gmu.edu/geo/ontology/owls/ap/v2/slope.owl
</Value>
</ValueList> </Slot>
...
</Service>

<Association ... associationType="urn:uuid:A80BE843-EB5E-43b1-
BBF6-90B90B8FA973" sourceObject="urn:uuid:19353DEB-CC81-44bb-
99D9-8A1F26FE5F41" targetObject="urn:uuid:26071546-50FE-494a-B2AB-
068B9E183F9A">
<Description><LocalizedString xml:lang="en-US" charset="UTF-8"
value="Slope Service associated InputDataType Ter-
rain_Elevation"/></Description>
...
</Association>

<Association ... associationType="urn:uuid:4EC1D47C-7D70-4b05-
A651-7AA8115FF831" sourceObject="urn:uuid:19353DEB-CC81-44bb-
99D9-8A1F26FE5F41" targetObject="urn:uuid:39DEB9D4-B134-4271-
8521-FCB8D3B142D9" >
<Description><LocalizedString xml:lang="en-US" charset="UTF-8"
value="Slope Service associated OutputDataType Ter-
rain_Slope"/></Description>
...
</Association>

<Classification ... id="urn:uuid:8F32D77D-C627-4d5f-9B30-
4926BD803A01" home="http://laits.gmu.edu:8099/csw/" ob-
jectType="urn:uuid:65e731a8-3325-4ac5-bd95-d71a277e3216"
status="Approved" classificationScheme="urn:uuid:011135B3-9EE1-439a-
ABCC-62A571B07175" classifiedObject="urn:uuid:19353DEB-CC81-44bb-
99D9-8A1F26FE5F41" classificationNode="urn:uuid:798B2817-4E67-402e-
A71A-A4D8C9E27360" nodeRepresentation="">
<Name><LocalizedString xml:lang="en-US" charset="UTF-8"
value="Slope Service Classification"/></Name>
<Description><LocalizedString xml:lang="en-US" charset="UTF-8"
value="This Slope Service is classified as a Slope Ser-
viceType"/></Description>
</Classification>
```

Fig. 11.8 Service semantics registraion

(5) The Service class is extended with a slot to specify the URI of an OWL-S file for each service instance. When a match is found based on the service capabilities, interaction with the service can be initiated through the specifications in the Process Model and Grounding parts of OWL-S. These interaction details are kept in the OWL-S file.

(6) Each Dataset, as an ebRIM CSWExtrinsicObject, is classified according to the "DataType" ClassificationScheme, using the associated ClassificationNode to specify its "DataType".

11.6 Semantic Search Functionalities

Combining the extensive research on UDDI with the research on ebRIM to provide a semantics-enabled search function in the ebRIM profile based implementation of geospatial catalogue service is useful. This study shows the architecture that supports a semantics-enabled search function, focusing on the application and extension of ebRIM in the geospatial domain. The first option from the research on UDDI (Sect. 11.3.1) is adopted. Since this option does not change the registry interface, it can help minimize the change on legacy systems. Figure 11.9 provides a simplified view of the interaction between the matching components and CSW when responding to a user's request.

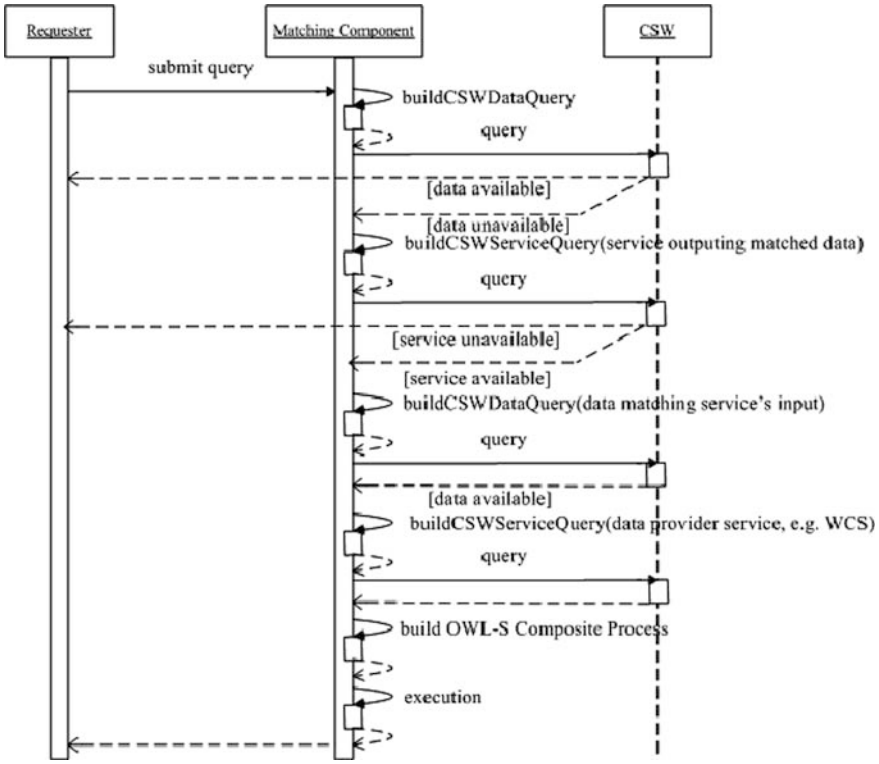

Fig. 11.9 UML sequence diagram illustrating the searching process (revised from Yue et al., 2007)

The semantic match function is based on the OWL knowledge base. The foundation of knowledge representation formalism for OWL is description logic (DL) (Baader and Nutt, 2003). The basic elements of description logic are concepts, roles, and constants. In the Web ontology context, they are commonly named classes, properties, and individuals respectively. Concepts group individuals into categories, roles stand for binary relations of those individuals and constants stand for individuals.

A DL knowledge base (KB) comprises two components: TBOX and ABOX. TBOX consists of a set of terminological axioms that make statements about how concepts or roles are related to each other. ABOX introduces individuals, i.e. instances of a class, into the knowledge base and gives the properties of these individuals. There are two types of reasoning in DL: TBOX reasoning and ABOX reasoning. In TBOX reasoning, a basic type of reasoning is to determine whether a concept is subsumed by another concept (i.e. subsumption reasoning). In ABOX reasoning, a basic type of reasoning is to determine whether a particular individual is an instance of a given concept description (i.e. instance checking).

Currently, semantic matching is based mainly on TBOX reasoning. Three types of match are defined: EXACT, SUBSUME, and RELAXED. Let OntR denotes the requested concept and OntP denotes the provider concept, the three matching conditions can be expressed in the following way, with decreasing priority order:

EXACT: OntR=OntP or OntR equivalent to OntP
SUBSUME: OntP is a subclassOf OntR
RELAXED: OntR is a subclassOf OntP

Users can set one of these matching conditions through the interface of the matching component.

There are three types of semantically augmented search functions:

(1) Dataset Search: The matching component gets semantically matched "DataType"s as the additional search condition in the standard CSW dataset query. For example, Figure 11.10 shows that in addition to the spatial and temporal filters, a "DataType" condition is added into the query of the "GetRecords" operation using the OGC filter specification.
(2) Service Search: It gets semantically-matched "ServiceType"s with the optional input/output "DataType"s as the additional search conditions in the standard CSW service query.
(3) "DataType"-Driven Service Chaining: A simplified process is illustrated in an UML sequence diagrams in Fig. 11.9. The composition is based on a match either between two services where the output of the first service provides the input of the second service, or between data and services, where the data provides the input of the service. Yue et al. (2007) have given a detailed description.

```
<csw:GetRecords xmlns="http://www.opengis.net/cat/csw"
xmlns:csw="http://www.opengis.net/cat/csw"
xmlns:ogc="http://www.opengis.net/ogc"
xmlns:gml="http://www.opengis.net/gml" version="1.0" responseHandler="" outputFormat="text/xml" charset="UTF-8" outputSchema="http://www.opengis.net/cat/csw" startPosition="1" maxRecords="50">
 <csw:distributedSearch propagate="ifNoLocal" hopCount="2" collectionId=""/>
 <csw:Query typeNames="WCSCoverage DataGranule Classification ClassificationScheme ClassificationNode">
 <csw:ElementSetName>full</csw:ElementSetName>
 <csw:ElementName>/WCSCoverage/</csw:ElementName>
 <csw:Constraint version="1.0.0">
 <ogc:Filter><ogc:And>
 <ogc:PropertyIsEqualTo>
<ogc:PropertyName>/WCSCoverage/granule</ogc:PropertyName>
 </ogc:PropertyIsEqualTo><ogc:PropertyIsLessThanOrEqualTo>
 <ogc:PropertyName>/DataGranule/beginDateTime</ogc:PropertyName>
  <ogc:Literal>2005-01-10T00:00:00Z</ogc:Literal>
 </ogc:PropertyIsLessThanOrEqualTo><ogc:PropertyIsGreaterThanOrEqualTo>
 <ogc:PropertyName>/DataGranule/endDateTime</ogc:PropertyName>
  <ogc:Literal>2005-01-10T23:59:59Z</ogc:Literal>
 </ogc:PropertyIsGreaterThanOrEqualTo>
 <ogc:BBOX>
 <ogc:PropertyName>/DataGranule/BBOX</ogc:PropertyName>
 <gml:Box srsName="EPSG:4326">
 <gml:coordinates>-122.2167,37.7994 -122.2167,37.7994</gml:coordinates>
 </gml:Box></ogc:BBOX>
 <ogc:PropertyIsEqualTo>
 <ogc:PropertyName>/WCSCoverage/@id</ogc:PropertyName>
 <ogc:PropertyName>/Classification/@classifiedObject</ogc:PropertyName>
 </ogc:PropertyIsEqualTo><ogc:PropertyIsEqualTo>
 <ogc:PropertyName>/Classification/@classificationScheme</ogc:PropertyName>
 <ogc:PropertyName>/ClassificationScheme/@id</ogc:PropertyName>
 </ogc:PropertyIsEqualTo><ogc:PropertyIsEqualTo>
 <ogc:PropertyName>/ClassificationScheme/Description/LocalizedString/@value</ogc:PropertyName>
 <ogc:Literal>http://www.laits.gmu.edu/geo/ontology/domain/GeoDataType.owl</ogc:Literal>
 </ogc:PropertyIsEqualTo><ogc:PropertyIsEqualTo>
 <ogc:PropertyName>/Classification/@classificationNode</ogc:PropertyName>
 <ogc:PropertyName>/ClassificationNode/@id</ogc:PropertyName>
 </ogc:PropertyIsEqualTo><ogc:PropertyIsEqualTo>
 <ogc:PropertyName>/ClassificationNode/@code</ogc:PropertyName>
 <ogc:Literal>Terrain_Elevation</ogc:Literal>
 </ogc:PropertyIsEqualTo>
 </ogc:And>
 </ogc:Filter></csw:Constraint></csw:Query>
 </csw:GetRecords>
```

Fig. 11.10 Dataset search

11.7 Conclusions

This chapter demonstrates how semantic search capability can be included in the ebRIM-based OGC CSW. The work explores the semantic representation of geospatial data and services to use the semantic relationship defined in OWL/OWL-S for semantic search in CSW. These semantics are organized in CSW through extending ebRIM elements. The semantic matching execution component that is implemented supports semantically augmented search for data and services. Such a semantically-augmented CSW can support "DataType"-driven service chaining with which a "Virtual Data Product" can be constructed and cataloged and such a product can be instantiated on-demand.

Acknowledgements This work is supported by US National Geospatial-Intelligence Agency NURI program (HM1582-04-1-2021), Project 40801153 supported by NSFC, National High Technology Research and Development Program of China (863 Program, No. 2007AA12Z214 and 2007AA120501), Specialized Research Fund for State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (Wuhan University), and Specialized Research Fund for State Key Laboratory of Software Engineering (Wuhan University).

References

Akkiraju, R., Goodwin, R., Doshi, P., Roeder, S., 2003. A Method for Semantically Enhancing the Service Discovery Capabilities of UDDI. In Proceedings of the Workshop on Information Integration on the Web, Eighteenth International Joint Conference on Artificial Intelligence (IJCAI), Mexico, pp. 87–92.

Baader, F. and Nutt, W., 2003. Basic Description Logics. In Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P., (eds.) The Description Logic Handbook. Theory, Implementation and Applications. Cambridge, Cambridge University Press, pp. 43–95.

Colgrave, J., Akkiraju, R., Goodwin, R., 2004. External Matching in UDDI. IEEE International Conference on Web Services, San Diego, USA, 8 pp.

Dean, M. and Schreiber, G., 2004. OWL Web Ontology Language Reference, W3C. http://www.w3.org/TR/owl-ref.

Dogac, A., Kabak, Y., Laleci, G. B., 2004. Enriching ebXML registries with OWL ontologies for efficient service discovery, In: Proceedings of the 14th International Workshop on Research Issues on Data Engineering: Web Services for E-Commerce and E-Government Applications (RIDE' 04), Boston, USA, pp. 69–76.

Dogac, A., Kabak, Y., Laleci, G. B., Mattocks, C., Najmi, F., Pollock, J., 2005. Enhancing ebXML registries to make them OWL aware, Distributed and Parallel Databases Journal, Springer-Verlag, July, Vol. 18, No. 1, pp. 9–36

Gruber, T. R., 1993. A translation approach to portable ontology specification, Knowledge Acquisition 5(2), pp. 199–220.

ISO/TC 211, 2003. ISO19115:2003, Geographic Information – Metadata.

ISO/TC 211, 2005. ISO19119:2005, Geographic Information – Services.

Martell, R., 2004. OGC™ Catalogue Services–ebRIM(ISO/TS 15000-3) profile of CSW, Version 0.9.1, OGC 04-017r1, Open Geospatial Consortium, Inc., 87 pp.

Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., Narayanan, S., Paolucci, M., Parsia, B., Payne, T., Sirin, E., Srinivasan, N., Sycara, K., 2004. OWL-based Web Service Ontology (OWL-S). http://www.daml.org/services/owl-s/1.1.

Nebert, D. and Whiteside, A., 2005. OGC™ Catalog Services Specification, Version 2.0.0, OGC 04-021r3, Open GIS Consortium Inc. 205 pp.

OASIS, 2004. Introduction to UDDI: Important Features and Functional Concepts. OASIS. 11 pp. http://uddi.org/pubs/uddi-tech-wp.pdf.

OASIS, 2005. ebXML Registry Information Model Version 3.0. OASIS Standard, 2 May, 2005, 78 pp.

Paolucci, M., Kawamura, T., Payne, T. R., Sycara, K., 2002a. Importing the Semantic Web in UDDI. In Web Services, E-Business and Semantic Web Workshop 2002, 12 pp.

Paolucci, M., Kawamura, T., Payne, T. R., Sycara, K., 2002b. Semantic Matching of Web Services Capabilities. Horrocks, I., Hendler, J. A. (eds.) The Semantic Web-ISWC 2002, First International Semantic Web Conference, Sardinia, Italy, June 9–12, 2002, Proceedings. Lecture Notes in Computer Science 2342, Springer, Berlin, Germany, 2002, 333–347.

Sivashanmugam, K., Verma, K., Sheth, A. P., Miller, J. A., 2003. Adding semantics to web services standards. In 1st Proceedings of the International Conference on Web Services, ICWS'03. Las Vegas, Nevada 2003, USA, 7 pp.

Srinivasan, N., Paolucci, M., Sycara, K., 2004. Adding OWL-S to UDDI, implementation and throughput. First International Workshop on Semantic Web Services and Web Process Composition, San Diego, USA 2004, 12 pp.

Vretanos, P. A., 2005. OpenGIS$^{@}$ filter encoding implementation specification. Version 1.1.0, OGC 04-095, Open Geospatial Consortium, Inc., 40 pp.

Wei, L., Keqing, H., Wudong, L., 2005a. Design and realization of ebXML registry classification model based on ontology. In: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05), pp. 809–814.

Wei, Y., Di, L., Zhao, B., Liao, G., Chen, A., Bai, Y., Liu, Y., 2005b. The design and implementation of a grid-enabled catalogue service. In: 25th Anniversary of IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2005), July 25–29, COEX, Seoul, Korea, pp. 4224–4227.

Yue, P., Di, L., Yang, W., Yu, G., Zhao, P., 2007. Semantics-based automatic composition of geospatial Web services chains. Computers & Geosciences, Vol. 33, No. 5, May, 2007, pp. 649–665.

Chapter 12
Geospatial Knowledge Discovery Using Semantic Web Services

Peisheng Zhao and Liping Di

12.1 Introduction

Modern-day satellites and other data acquisition systems have collected an overwhelming volume of Earth and space science data. The data are processed and managed by a variety of geographically distributed data providers. NASA's Earth Observing System (EOS), for instance, has been generating on the average almost 100 gigabytes of imagery per hour for the past decade. It releases over 900 Earth science data products at more than a dozen data centers. It is extremely valuable for innovative scientific researches and decision-making processes to extract useful information and knowledge from these distributed massive volumes of data. A geospatial model in which prior domain expertise is encoded formally as computable algorithms can facilitate knowledge discovery by detecting and interpreting patterns and regularities, discovering classification rules, and inferring causation. With complex spatial and/or temporal dynamics, geospatial knowledge discovery commonly requires a capability beyond that of an individual geospatial model. Specifically, it involves a complex workflow that requires the integration of various geospatial models and distributed multi-disciplinary, multi-source, and multi-scale science data. For example, to predict fire behavior and estimate possible damage, decision makers and firefighters must effectively combine satellite observations, weather data, geographic data, census data, and simulation models from various sources. The best model and the most appropriate data must be selected in order to predict the fire spread in near-real time or real time. Therefore, the interoperability of geospatial models and data is becoming a critical issue for geospatial knowledge discovery.

P. Zhao (✉)
Center for Spatial Information Science and Systems (CSISS), George Mason University, Greenbelt, MD 20770, USA
e-mail: pzhao@gmu.edu

L. Di, H.K. Ramapriyan (eds.), *Standard-Based Data and Information Systems for Earth Observation*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-540-88264-0_12,

The number and amount of geospatial data and models available online as Web services has been increasing in recent years. This increase significantly enhances the ability of users to collect and analyze geospatial data from different systems interoperably. The NASA Earth Science Gateway (http://esg.gsfc.nasa.gov) allows users streamlined access to scientific and research products, including data, models, and visualizations provided by a variety of national and international organizations, through open standard Web protocols. The GeoBrain Processing Web Services (http://geobrain.laits.gmu.edu:81/grassweb/manuals/) based on Geographic Resource Analysis Support System (GRASS) functionality modules provides the capabilities of geospatial data management, raster image processing, spatial modeling and analysis, graphic map production, and visualization over the network. The Adam Web Services (http://datamining.itsc.uah.edu/adam/) leverages the Algorithm Development and Mining toolkit to enable mining remotely sensed and other scientific data dynamically over the network for pattern recognition, image processing and optimization, and association rule exploration. All these achievements in interoperable geospatial Web services make discovery of geospatial knowledge in a Web service environment possible. However, such methods may require the user to understand more about the Web services for the variety of geospatial models and have better computational skills than their training provides. The expected increase in geospatial Web services will exacerbate the difficulties in service discovery, integration and reuse for different applications because of the semantic gap resulting from ambiguous descriptions of service properties, capabilities, interfaces, and effects. For example, today's retrieval methods are typically limited to word (string) match, and do not exploit the meaning of the underlying objects. Because of the lack of semantics in Web Services, implementing reliable and large-scale interoperation of computer programs or agents is impossible (McIlraith et al. 2001). Therefore, a way to characterize geospatial models and relevant datasets for the purpose of easy discovery and accessibility, specifically for efficient machine-accessible, is needed to automatically or semi-automatically discover geospatial knowledge. The Semantic Web provides a promising common interoperable framework in which information is given a well-defined meaning in an unambiguous and computer-interpretable form by using ontology. Data and services can then be used for more effective discovery, automation, integration, and reuse in various applications. This paper proposes a new approach to geospatial knowledge discovery using semantic Web services. This approach provides a framework that makes individual data sets and geospatial models discoverable and accessible. It also makes them interoperable, allowing easy assembly into workflows that implement geospatial knowledge discovery.

The reminder of this paper is organized as follows. Section 12.2 discusses geospatial Web services and their application to geospatial knowledge discovery. In Sect. 12.3, geospatial ontology and the semantic Web are discussed. The approach architecture and technical detail are described in Sect. 12.4. And finally, Sect. 12.5 presents conclusions and plans for future work.

12.2 Geospatial Knowledge Discovery and Web Services

Generally, for geospatial knowledge discovery guided by domain-specific expertise has the following components, as shown in Fig. 12.1:

- Data selection, in which geospatial data is discovered and retrieved from personal databases/file archives, distributed data centers, and live field data.
- Data pre-processing, in which geospatial data is represented in different ways, such as by spatial-temporal subsetting, coordinate transformation, feature reduction, and format translation.
- Data mining, in which geospatial models are used to extract information and knowledge from geospatial data, such as by classification, prediction, clustering, detection, and association.
- Data visualization, in which geospatial knowledge is represented through graphical means, such as thematic maps, statistical charts, and map overlays.

Geospatial knowledge discovery involves a complex workflow usually requiring integration of geospatial models and data with control structures such as sequence, parallel, switch and while. Such a workflow probably contains multiple steps of distributed computation for mining multi-disciplinary, multi-source, and multi-scale scientific data. Therefore, the performance of geospatial knowledge discovery highly depends on data and model accessibility and their integrations over a network.

The advance of Service-Oriented Architecture (SOA) which is consisted of a collection of Web services promise standard-based information interoperability. A Web service is "a software system designed to support interoperable machine-to-machine interaction over a network (http://www.w3.org/TR/ws-arch). It is capable of collaborating process control and sharing data and information across many applications and different platforms. It has the following novel features:

- Discrete functionalities encapsulated as a loosely coupled, reusable distributed component.
- A standard interface described in machine-processable Web Service Description Language (WSDL) to hide all the details of implementation.
- Programmatic access by standard Internet protocols using Simple Object Access Protocol (SOAP) messages.
- Assembly of individual Web Services into a service chain to complete a more complicated task.

In recent years, a growing number of geospatial Web services designed to deal with distributed geospatial data have emerged as Web services technologies mature. The term "geospatial Web service" is straightforward insofar as it refers to using Web service technologies to manage, analyze, and distribute geospatial data. But it

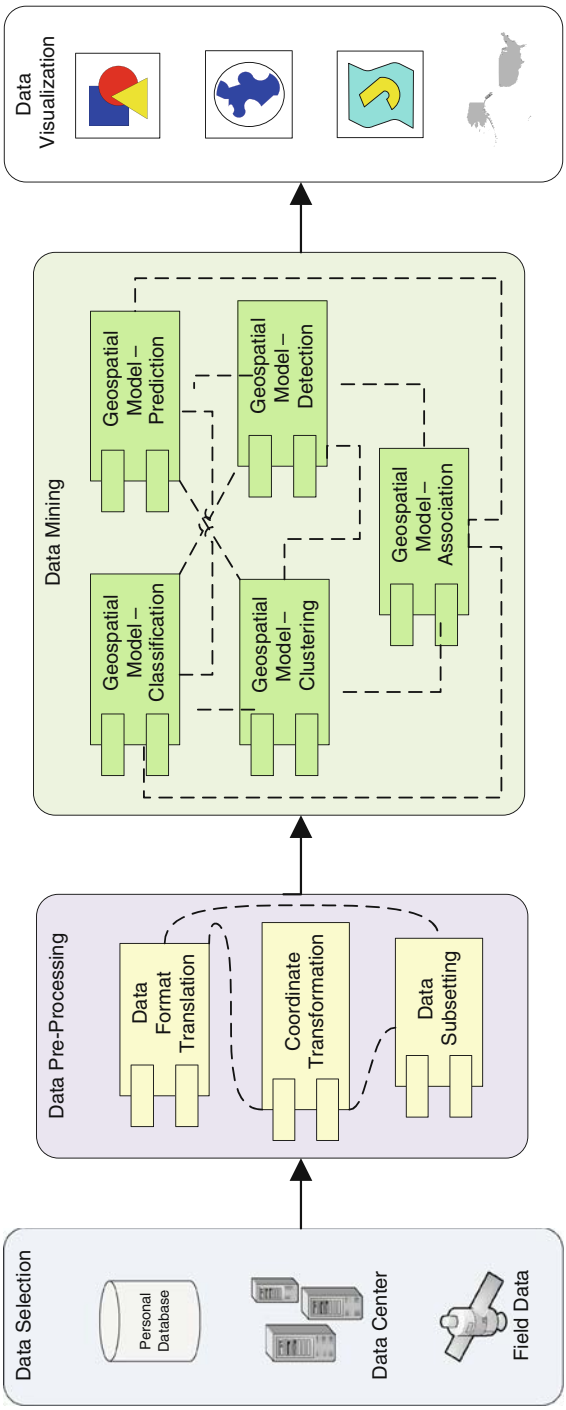

Fig. 12.1 Geospatial knowledge discovery

further refers to the architecture, standards, and patterns that make geospatial Web services feasible. Geospatial Web services are changing the way in which geospatial information systems and applications are designed, developed and deployed (Zhao et al. 2006). According to the Open Geospatial Consortium (OGC) specifications, which make complex geospatial information and services accessible and interoperable with all kinds of applications, geospatial Web services can be categorized from the prospect of geospatial knowledge discovery, as shown in Fig. 12.2:

- *Geospatial Data Service*: By having well-defined interfaces with explicit syntax, the services in this category allow the user to (1) retrieve distributed heterogeneous geospatial data, using services such as OGC Web Coverage Service (WCS) (Whiteside and Evans 2008), OGC Web Feature Service (WFS) (Vretanos 2005), and OGC (Sensor Observation Service (SOS) (Na and Priest, 2007), (2) plan and harvest geospatial data, for example, by using the OGC Sensor Planning Service (SPS) (Simonis, 2005), OGC WCS Transactional (WCS-T), and OGC WFS Transactional (WFS-T), and (3) pre-process geospatial data by using the functions embedded in WCS and WFS, such as subsetting, resampling, format translation and coordinate transformation, in order to make the data more useful for the desired purpose.
- *Geospatial Processing Service*: Its definition is similar to the definition of Web Processing Service (WPS) in (Schut 2007). The Geospatial Processing Service provides client access to pre-programmed calculations and computational models that operate on spatial reference data for pattern recognition, feature processing, optimization, and association rule exploration. It actually changes the physical meaning (type) of geospatial data. OGC WPS provides a set of standard interfaces to represent and access to the Geospatial Processing Service.
- *Geospatial Visualization Service*: To effectively represent abstract and concrete geospatial data, information and knowledge, a Geospatial Visualization Service creates multidimensional images, diagrams, and animations as seen from multiple prospects. For example, OGC Web Map Service (WMS) (Beaujardiere, 2006) supports the networked interchange of geospatial information dynamically from real geographical data as a "map" which is generally rendered in a spatially referenced pictorial image, using pre-defined formats, for example, PNG, GIF or JPEG.
- *Geospatial Facility Service*: Service support is the essential part of the various components that support an SOA solution. Geospatial Facility Services include a set of services that facilitate service management and service orchestration. For example, OGC Geospatial Digital Rights Management (GeoDRM) (Vowles, 2006) implements access authorization and authentication by assigning digital rights to distributed geospatial resources, The Workflow Management Service empowers users to compose, manage and execute more complicated geospatial service chains. The OGC Sensor Alert Service (SAS) and Web Notification Service (WNS) enable the user to subscribe to and receive messages or alerts from sensor services or other elements of service workflows.

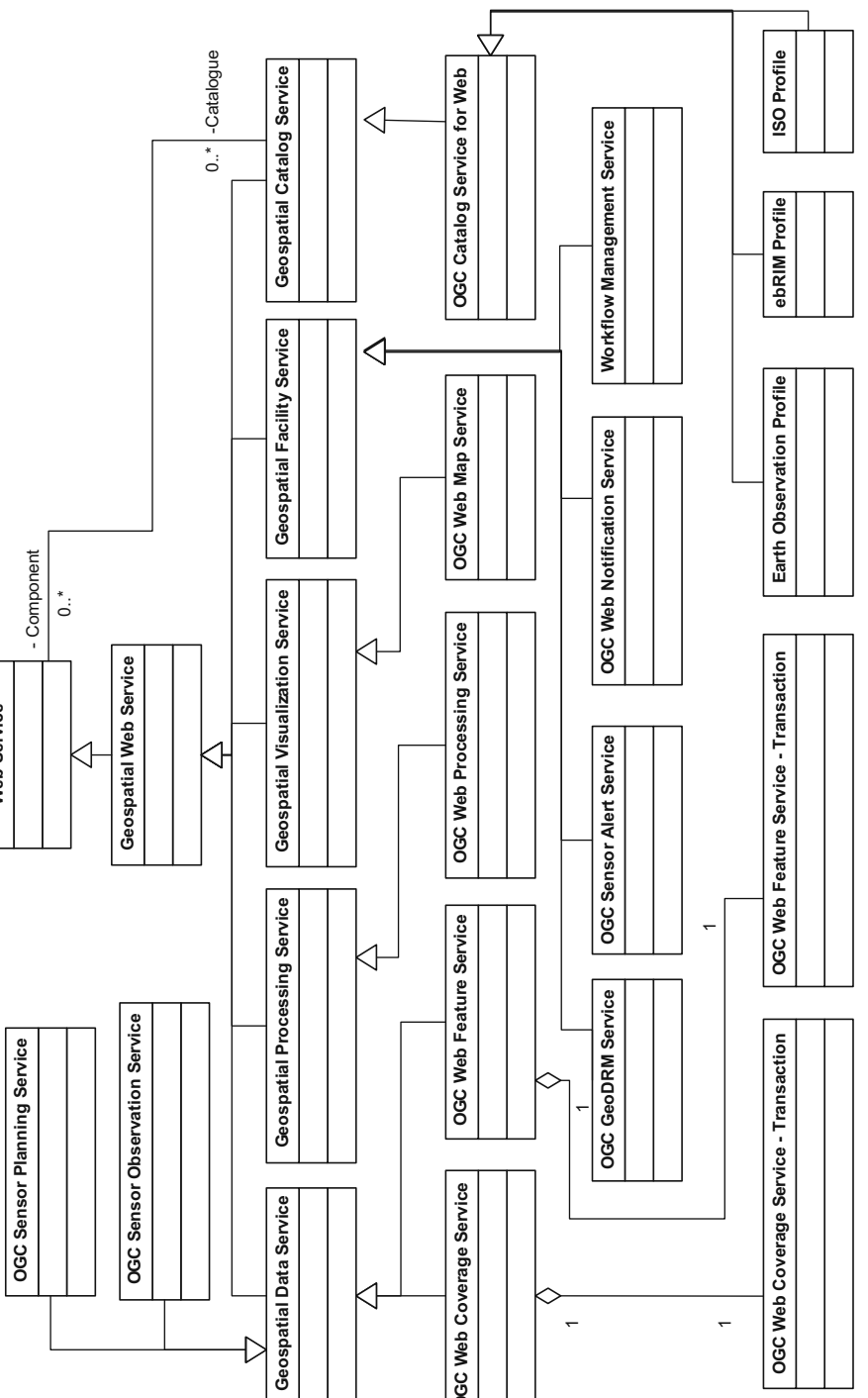

Fig. 12.2 Geospatial Web services

- *Geospatial Catalog Service*: An important part of the SOA approach is the extensive use of metadata in service and data directories. The Geospatial Catalog Service provides information about the functional capabilities and operational characteristics of services and the spatial-temporal characteristics of data. It acts as a directory, allowing providers to advertise geospatial resources using standard meta-information, and requestors to discover the geospatial resources they need by querying meta-information. OGC Catalog Service for Web (CS-W) (Nebert et al. 2007) is becoming the *de facto* standard for geospatial SOA. It specifies a framework for defining the application profiles required to publish and access digital catalogues of metadata for geographic data, services, and related resource information, for example, the Electronic Business Registry Information Model (ebRIM) profile, the International Organization for Standardization (ISO) metadata profile, and the Earth Observation (EO) product profile.

With the proliferation of geospatial Web services, it is desirable to use a service chain to implement the complicated workflow of geospatial knowledge discovery. Assembling individual services into a service chain is called service orchestration. Because most geospatial Web services are aimed at interoperability, especially those services based on domain standards, service orchestration has great potential for reducing the effort involved in geospatial knowledge discovery. A geospatial Web service can be orchestrated in three modes, as stated in the definition of a service chain in ISO 19119 (2005):

- *Transparent*: The user knows everything about geospatial knowledge discovery and plays a central role in finding and composing all the required services and data. The composite service chain can be invoked either in a user-controlled sequence or in a system-controlled process.
- *Translucent*: The user queries the system for each step of geospatial knowledge discovery, and then the system assists the user to select and configure the most suitable services and data for building a service chain.
- *Opaque*: The user presents a problem, and then the system uses its embedded knowledge to automatically build a service chain using the best services and data, without the user's intervention.

The key distinctions between these patterns are the differences in visibility to the user of the data and services and the level of expertise and control required. In the transparent pattern, control is exclusively with the user requiring a domain expert who is sufficiently knowledgeable about not only the domain model and data, but also the low-level technical detail of service composition. In the opaque pattern, control is exclusively with the powerful system. The domain expert need not know anything about services and data. It is most like a "black box", and sometimes, because the user is not involved, it does not reflect the real purpose of the domain expert. In the translucent pattern, expertise and control are shared by the system and the user. This two-way interaction allows the user to address only optimizing the workflow for geospatial knowledge discovery and allows the system to address

only the necessary technical detail about service composition. Thus, it reduces the complexity of the work required by both the system and the user and improves the efficiency of geospatial knowledge discovery. This paper focuses on illustrating the translucent approach whose core problem is to discover which services and data are most appropriate for the user requests.

12.3 Ontologies for Geospatial Knowledge Discovery

One of the most important tasks for geospatial knowledge discovery in the approach proposed here is to find the geospatial processing services and geospatial data services that match specific requirements. Metadata is always used in describing and discovering geospatial Web services. However, mismatching may arise due to one of the following factors: (Zhao et al. 2006):

- *Lack of semantics*: Generally, service outputs are represented by basic data types, for example, string, integer, and double. If the match criteria are based strictly on data type, the results returned may contain numerous improper matches. For example, a URI (Universal Resource Identification) string may point to either an actual data location or a data schema location.
- *Incomplete semantics*: Even if the inputs and outputs of services are semantically matched, the results returned may not satisfy the intended requirement. This is because an improper computation algorithm may have been used. For example, there are many algorithms for slope calculations, for example, the neighborhood method, the quadratic surface method, the maximum slope method, and the maximum downhill method. If the quadratic surface method is desired, matches with other methods are an error for this search task.
- *Lack of relationship semantics*: It is common knowledge that service A can be used in place of service B if service A includes the contents of service B. For example, a land cover map can be used in place of a vegetation map, since the vegetation categories (forest land, grassland, and cultivated land) are available in the land cover map. Without a clear definition of the relationship between A and B in the semantics of the metadata, the search would not be able to make this intelligent association.

The advent of the geospatial semantic Web enables capture of the semantic network of the geospatial world. It allows intelligent applications to take advantage of build-in geospatial reasoning capabilities to build service chains for deriving knowledge (Zhao et al. 2007, Egenhofer 2002, Lieberman et al. 2005). As a critical component of the Semantic Web, ontology provides a common understanding of domain knowledge in a generic way for sharing across applications and groups (Chandrasekaran et al. 1999). Geospatial ontologies provide the following functions for geospatial knowledge discovery:

- *Semantic interoperability*: Geospatial ontologies define geospatial terms in a formal method. This method sufficiently captures the semantic details of geospatial

concepts to provide a common understanding of geospatial data and processes. With geospatial ontologies, geospatial Web services are well-defined and can be chained together without semantic ambiguity.
- *Geospatial reasoning*: By defining geospatial associations and patterns, geospatial ontologies make it possible to infer geospatial concept relationships and computational plans. Therefore, geospatial Web services are well-associated by their algorithm descriptions and external interfaces (inputs and outputs).
- *Reuse and organization of information*: Geospatial ontologies enable standardization of libraries or repositories of geospatial information. The workflows for geospatial knowledge discovery are thus easily reused without losing semantics.

To facilitate knowledge discovery, more and more researches have been focusing on using ontology to encapsulate the semantics of data mining techniques and relates them to the concepts within disciplines recently. In (Cannataro and Comito 2003), the data mining ontology is presented in order to simplify the development of distributed knowledge discovery. This ontology offers a reference model for different kinds of data mining tasks, methodologies and software, which helps users to find the most appropriate mining solutions. However, it is oriented to general data mining problems, so that it conceptualizes only generic data mining techniques and does not consider domain concepts and data. As scientific problems are more complex, that relates the data to the geospatial domain concepts by using ontology is proposed in (Hwang 2004). Users can thus easily retrieve the relevant data sets to be compared by navigating the ontology. However, that work is not concerned with data mining techniques. A conceptual framework which exploits ontology to describe the domain model and task model of geospatial data mining algorithms is described in (Durbha and King 2004). Furthermore, a framework based on concept model and domain-dependent ontologies is discussed in (Raskin 2004). In that framework, the basic domain concepts are identified first and later generalized depending upon the level of reasoning required for executing a particular query. However, neither of those two frameworks considers computational issues. Therefore, a set of comprehensive geospatial ontologies that cover domain concepts (model, task, and data) and computational details (services and workflow) are needed for geospatial knowledge discovery.

12.4 Framework for Semantic Web Service-based Geospatial Knowledge Discovery

12.4.1 Framework Architecture

The ultimate goal of this framework shown in Fig. 12.3 is to use semantic Web services to enable individual geospatial data and models not only discoverable and accessible, but also interoperable. They can then be easily assembled into workflows that implement the complex tasks required for geospatial knowledge discovery.

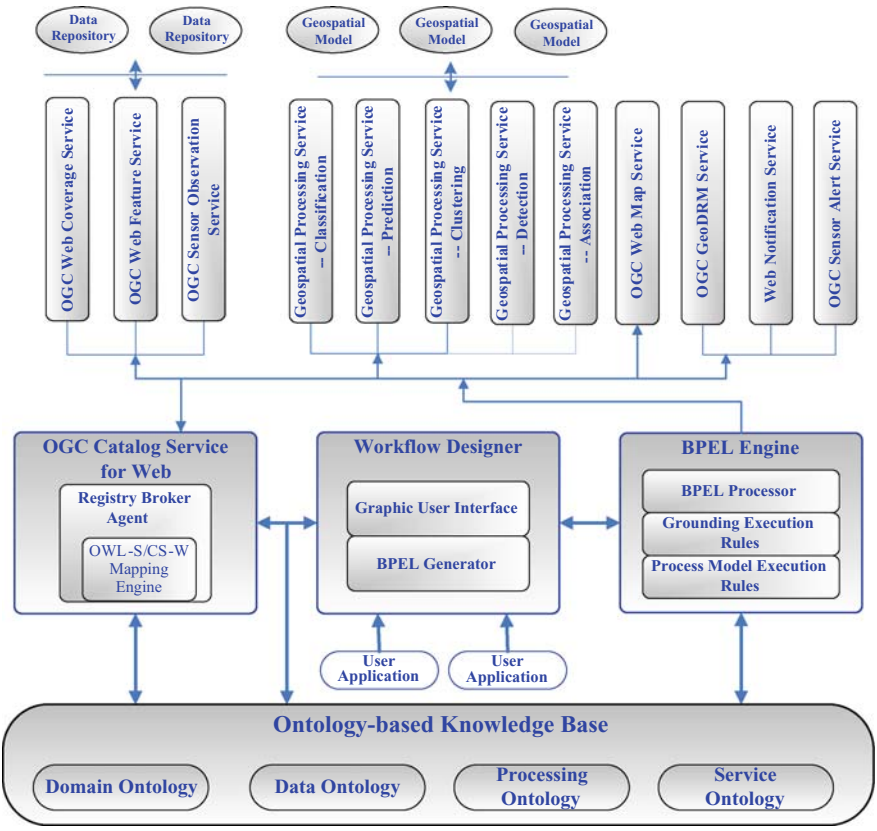

Fig. 12.3 Framework architecture

The framework covers Web services for geospatial data and models, catalogue services, workflow editors and engines, and ontology-based knowledge bases. The component functionalities are illustrated in the following sections.

12.4.2 Ontology-Based Knowledge Base

The knowledge base provides the knowledge on semantic matching of geospatial Web services. In this knowledge base, the set of hierarchical ontologies shown in Fig. 12.4 is implemented. These ontologies define the domain concepts and the linkage metrics for geospatial data, processing and scientific problems. Wherever possible, these ontologies are based on existing standards or agreements within the geospatial domain rather than being designed from scratch. The use of ontologies to describe geospatial processing and their relevant datasets gives a well-defined semantic meaning for the diverse data sources and the tasks to process them. Thus, an ontology-based knowledge base can help users to efficiently find the best solution and the most appropriate data for geospatial knowledge discovery.

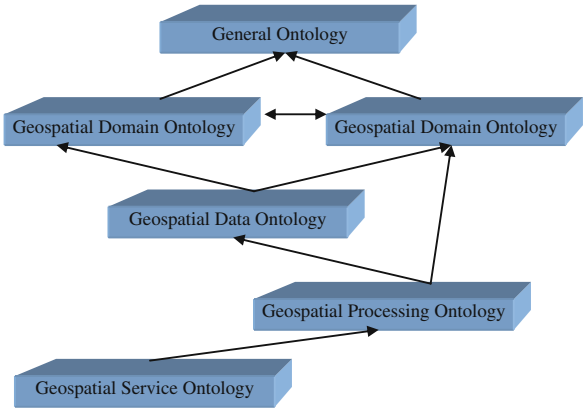

Fig. 12.4 Ontologies for geospatial knowledge discovery

The general ontology is the core upper level vocabulary for all human consensus concepts. It defines common terms to which all other ontologies refer. The Dublin Core Metadata (http://dublincore.org/) and OpenCyc (http://www.opencyc.org) are the bases of definitions of upper level concepts and assertions about these concepts.

Geospatial domain ontologies aim at providing the core conceptualization and knowledge structure of the variety of geospatial domains. For example, "Erosion Sedimentation" belongs to the domain of "Land Surface", and "Landslide" is a kind of "Erosion Sedimentation". In the knowledge base, the domain ontology represents the problem space over which the user will query. Other ontologies directly or indirectly incorporate geospatial domain ontologies. Recently, several projects have developed ontologies across different geospatial domains. The ontologies implemented in the Semantic Web for Earth and Environmental Terminology (SWEET) (http://sweet.jpl.nasa.gov/ontology/), described in the Web Ontology Language (OWL), contain several thousand terms covering a broad extent of Earth system science. SWEET is used as a starting point. Its contents are reorganized and expanded to cover additional geospatial domain concepts by incorporating the terms in the Global Change Master Directory (GCMD) (http://gcmd.nasa.gov/), the Earth Science Modeling Framework (ESMF) (http://www.esmf.ucar.edu/), and the General Multilingual Environmental Thesaurus (GEMET) (http://www.eionet.europa.eu/gemet). Geospatial domain ontologies cover the semantics of (1) spatial-temporal factors, e.g. location, time and units, (2) physical facts, e.g. physical phenomena, physical properties and physical substances, (3) disciplines, e.g. scientific domains and projects, and (4) data collection, e.g. instruments, platforms and sensors.

Geospatial data ontology provides an ontological view of distributed heterogeneous geospatial data resources. It describes data identification, quality, organization, spatial reference, entities and attributes, and distribution. In addition, it directly incorporates geospatial domain ontologies to link the data with scientific research areas and real physical facts to ensure that geospatial data more discoverable, usable and interoperable, as shown by the following code:

```
<owl:Class rdf:ID="Physical_Fact"/>
<owl:Class rdf:ID="Domain"/>
<owl:Class rdf:ID="Data_Type"/>
<owl:ObjectProperty rdf:ID="belongTo">
  <rdfs:range rdf:resource="#Domain"/>
  <rdfs:domain rdf:resource="#Data_Type"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="usedFor">
  <rdfs:range rdf:resource="#Physical_Fact"
  <rdfs:domain rdf:resource="#Data_Type"/>
</owl:ObjectProperty>
```

ISO 19115 and the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata are the widely used metadata standards for geospatial metadata. NASA uses the metadata of Earth Observing System Data and Information System (EOSDIS) Core System (ECS) to describe geospatial data distributed in NASA data centers. From the perspective of operation, geospatial data ontology adds semantics to the metadata. Users can locate data without knowing the exact metadata keywords, because the query terms may have an equivalent definition in the geospatial domain ontology. To provide a unified view of metadata, the semantic relationships among the terms in different metadata standards are defined here, such as subclass and same as. Thus, there is no distinct boundary between any two metadata standards. The user can use a term from any of the metadata standards to query the data described in any one of the other metadata standards.

Geospatial processing ontology provides a reference model for different kinds of geospatial models. It directly incorporates the geospatial domain ontologies and geospatial data ontology to associate geospatial models with scientific problems and relevant data sources. This ontology represents the features of the available geospatial models, classifies their internal structures, and documents the relationships and the constraints among them. The following important concepts related to geospatial models are included in this ontology:

```
<owl:Class rdf:ID="Geo_Processing"/>
<owl:Class rdf:ID="Data_Source"/>
<owl:Class rdf:ID="Mission"/>
<owl:Class rdf:ID="Quality_Measurement"/>
<owl:Class rdf:ID="Methodology"/>
<owl:Class rdf:ID="Result"/>
<owl:ObjectProperty rdf:ID="quality">
  <rdfs:range rdf:resource="#Quality_Measurement"/>
  <rdfs:domain rdf:resource="#Geo_Processing"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="isFor">
  <rdfs:domain rdf:resource="#Mission"/>
  <rdfs:domain rdf:resource="#Geo_Processing"/>
```

```
</owl:ObjectProperty>
<owl:FunctionalProperty rdf:ID="input">
  <rdfs:range rdf:resource="#Data_Source"/>
  <rdfs:domain rdf:resource="#Geo_Processing"/>
</owl:FunctionalProperty>
<owl:FunctionalProperty rdf:ID="output">
  <rdfs:range rdf:resource="#Result"/>
  <rdfs:domain rdf:resource="#Geo_Processing"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="uses">
  <rdfs:range rdf:resource="#Methodology"/>
  <rdfs:domain rdf:resource="#Geo_Processing"/>
</owl:ObjectProperty>
```

- Mission: the specific goal of a geospatial model, such as classification or clustering.
- Methodology: the methodology or algorithms used in the geospatial model, such as neural network and Bayesian network.
- Data_Source: the type of data and the sources on which the geospatial model can work. The data type is defined in data source ontology.
- Result: the properties of output of a geospatial model.
- Quality: the quality measures of a geospatial model, such as running time and accuracy.

Geospatial service ontology enables semantic matching to discover, invoke, and compose services. It directly incorporates geospatial processing ontology to describe geospatial Web services in OWL-based Web Service Ontology (OWL-S). The profile of OWL-S, which describes who provides the service, what the service does, and other properties of services, allows the knowledgebase to infer whether or not a particular service is appropriate to a given problem. The OWL-S process model, which states the inputs, outputs, preconditions and effects of a service, allows the knowledgebase to determine whether or not a service meets the intended requirements as well as the conditions to invoke the service. The OWL-S grounding, which presents the ports, protocols and encoding of invocation, tells how to invoke a service.

12.4.3 Building Workflow

Figure 12.5 illustrates a simplified sequence view of how a workflow for geospatial knowledge discovery is built.

By incorporating the ontologies in the knowledge base, the OGC CS-W in the framework supports flexible semantic matching regardless of syntactic differences, especially the “exact”, “plug in” and “subsume” matching of:

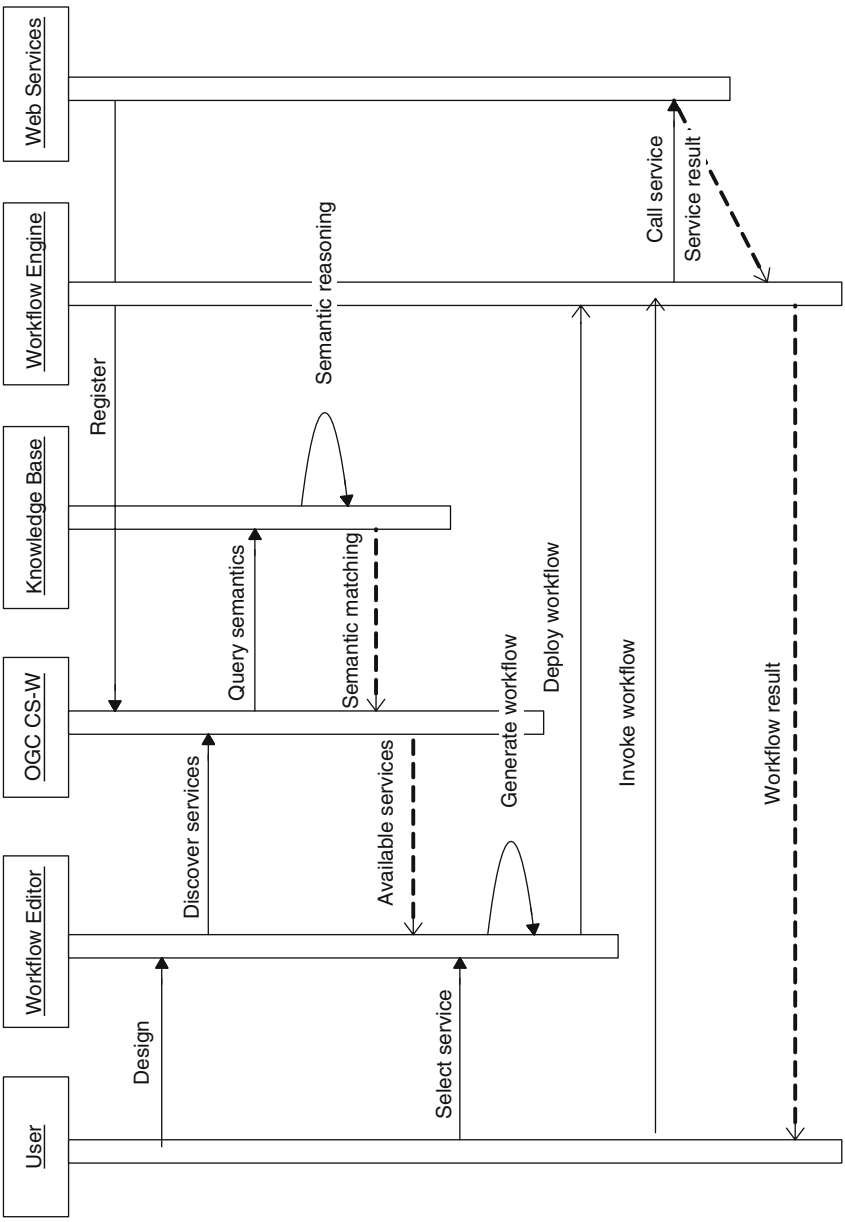

Fig. 12.5 Sequence diagram of geospatial knowledge discovery

- *Dataset Search*: Semantically matched data types are the additional search condition in the standard CS-W dataset query.
- *Service Search*: Semantically matched service types with optional associated data types are the additional search conditions in the standard CS-W service query.

In the CS-W, each geospatial service is associated with a service type and has been related to preconditions, and each data item is associated with a data type. Therefore, geospatial service discovery can be formalized as service type matching, data type matching and condition matching (Zhao 2009):

$$Match_{plug\text{-}in\text{-}dataType}(DS, DQ) = D_{DS} \subseteq D_{DQ} \wedge PF_{DS} \subseteq PF_{DQ} \quad (1)$$

$$Match_{plug\text{-}in\text{-}serviceType}(SS, SQ) = M_{SS} \subseteq M_{SQ} \wedge I_{SQ} \subseteq I_{SS} \wedge O_{SS} \subseteq OSQ \quad (2)$$

$$Match_{Plug\text{-}in\text{-}Effect-Pre}(S, Q) = S_{pre} \Rightarrow Q_{effect} \quad (3)$$

A concept "*Data_Type*" DS is assumed to be a plug-in match for a requested "*Data_Type*" DQ (1) if all properties of the concept, including "*Discipline*" D_{DS}, and "*Physical_Fact*" PF_{DS}, subsume the counterparts of the DQ, i.e. D_{DQ}, PF_{DQ}. A concept "*Service_Type*" SS: (I_{SS}, M_{SS}) → O_{SS} is then assumed to be a plug-in match for a requested "*Service_Type*" SQ: (I_{SQ}, M_{SQ}) → O_{SQ} (2) if the "*Data_Type*" of the concept input I_{SS} is more generic and the "*Data_Type*" of the concept output O_{SS} is more specific, and the *Methodology* of the concept M_{SS} is more specific than of requested one. The *Plug-in Effect-Pre Match* holds when the preconditions of service *S* are more general than the effect of service *Q* (3). This match means that service *S* can be connected with service *Q* irrespective of the data representation. If no match is found, an additional service may have to be provided in the service chain to establish this match.

Figure 12.6 shows the Workflow Editor designed to allow users to build a workflow by drag-and-drop in collaboration with the CS-W. The left column shows the lists of data and services registered in the catalogue service. The right column is the

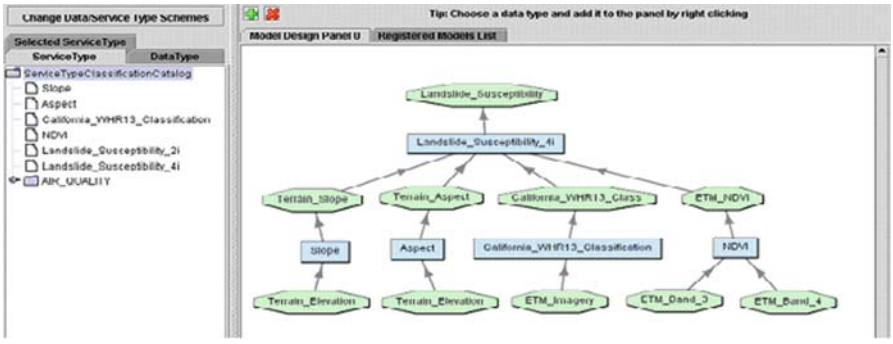

Fig. 12.6 Workflow editor

workflow graphic representation: the blue rectangles represent services, the green polygons represent data, and the black lines with arrows represent I/O control structures. To eliminate the possibility of match error between available components and the user request, only those services whose output is more specific than the goal state and whose input is more general than the user wants are listed and selectable. If more than one service satisfies these conditions, the editor allows the user to view the metadata of the services to assist selection.

The widely used Business Process Execution Language (BPEL) for Web Services (BPELWS) is used to encode workflows. A Workflow Engine, which is a BPEL process manager for integrating services into collaborative and transactional processes within the SOA, is designed and implemented to manage, deploy, and execute workflows. Figure 12.7 shows the engine's Web interface. WSDL-based Web services and BPEL-based Web service chains can be validated, deployed, and executed dynamically by this engine. The engine has evolved to one much more robust than many other contemporary engines. It can support complex schemas, e.g. substitution groups and multiple occurrences of elements. It can handle the schemas used by OGC, especially Geography Markup Language (GML). Different invocations, for example HTTP POST/GET and the Simple Object Access Protocol (SOAP), are supported. Moreover, this engine is deployed as a standard Web service so it can be easily integrated into other applications.

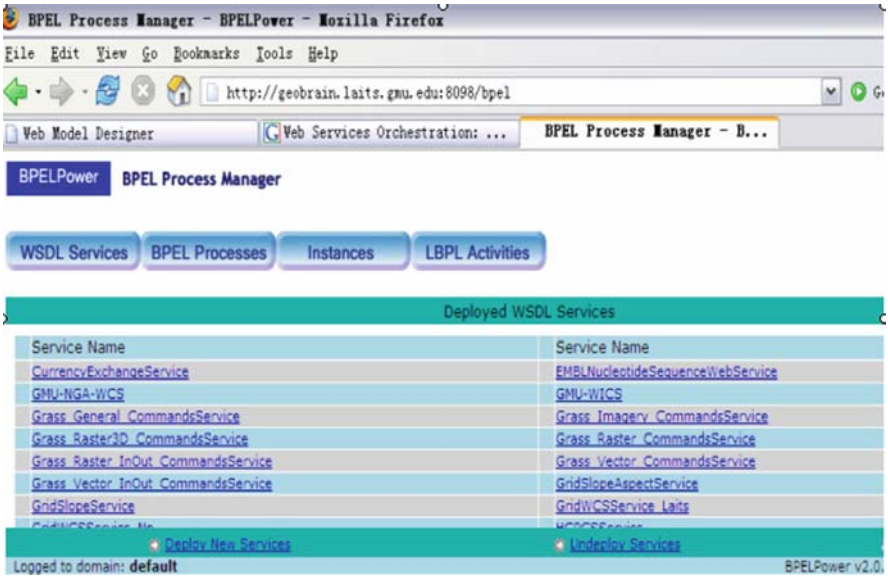

Fig. 12.7 Web interface of workflow engine

12.5 Conclusion

The most significant feature of the proposed approach is the use of semantic Web services to bridge the growing interoperability gap between data collection and analysis. This gap hinders geospatial knowledge discovery. In the proposed approach, domain concepts are well defined by geospatial ontology as the basic knowledge, and data and geospatial models are then well described by these concepts and served by OGC Web services and semantic Web services. The whole process of geospatial knowledge discovery is to build a workflow (service chain) in predefined patterns of domain concepts. This approach provides a mechanism by which scientists and decision-makers can fully exploit the potential of the geospatial data and models across domains.

Ontology plays a critical role in the proposed approach. However, it is impossible to exhaustively put all relevant geospatial information into ontologies. The next step is to investigate more existing geospatial ontologies and standards to sketch geospatial space precisely and elaborate the relationships inherent in the nature of geospatial data and data mining technologies.

References

Beaujardiere, J.: OpenGIS Web Map Server Implementation Specification OGC 06-042. Technical report, Open Geospatial Consortium Inc. (2006)

Cannataro, M. and Comito, C.: A Data Mining Ontology for Grid Programming. In: 1st International Workshop on Semantics in Peer-to-Peer and Grid Computing, pp. 113–134 (2003)

Chandrasekaran, B., Johnson, T. and Benjamins, V.: Ontologies: what are they? why do we need them? IEEE Intelligent Systems and Their Applications 14, 20–26 (1999)

Durbha, S. and King, R.: Knowledge mining in earth observation data archives: a domain ontology perspective. In: 2004 IEEE International Geoscience and Remote Sensing Symposium (2004)

Egenhofer, M.: Toward the semantic geospatial web. In: 10th ACM International Symposium on Advances in Geographic Information Systems, pp. 1–4 (2002)

Hwang, S.: Using formal ontology for integrated spatial data mining. In: International Conference on Computational Science and Its Applications, pp. 1026–1035, Springer-Verlag, New York (2004)

ISO/TC211: ISO 19119:2005 Geographic Information – Services. Technical report (2005)

Lieberman, J., Pehle, T. and Dean, M.: Semantic evolution of geospatial web services. In: W3C Workshop on Frameworks for Semantic in Web Services. The World Wide Web Consortium (W3C) (2005)

McIlraith, S., Son, T. and Zeng, H.: Semantic Web services. IEEE Intelligent Systems 16, 46–53 (2001)

Na, A. and Priest, M.: Sensor Observation Service OGC 06-009r6. Technical report, Open Geospatial Consortium Inc. (2007)

Nebert, D., Whiteside, A. and Vretanos, P.: OpenGIS Catalogue Services Specification OGC 07-006r1. Technical report (2007)

Raskin, R.: Enableing Semantic Interoperability for Earth Science Data. Technical report, NASA JPL (2004)

Schut, P.: OpenGIS Web Processing Service OGC 05-007r7. Technical report, Open Geospatial Consortium Inc. (2007)

Simonis, I.: OpenGIS Sensor Planning Service Implementation Specification OGC 07-014r3. Technical report, Open Geospatial Consortium Inc. (2005)
Vowles, G.: Geospatial Digital Rights Management Reference Model OGC 06-004r3. Technical report (2006)
Vretanos, P.: Web Feature Service Implementation Specification OGC 04-094. Technical report, Open Geospatial Consortium Inc. (2005)
Web Service Architecture, http://www.w3.org/TR/ws-arch
Whiteside, A. and Evans, J.: Web Coverage Service (WCS) Implementation Standard OGC 07-067r5. Technical report, Open Geospatial Consortium Inc. (2008)
Zhao, P., Di, L., Yang, W., Yu, G. and Yue, P.: Geospatial semantic web: critical issues. In: H. Karimi (eds.) Encyclopedia of Geoinformatics. Idea Group Publishing, Hershey (2007)
Zhao, P., Di, L., Yue, P., Yu, G. and Yang, W.: Semantic web based geospatial knowledge transformation. Computer & Geosciences 35(4), 798–808 (2009)
Zhao, P., Yu, G. and Di, L.: Geospatial web services. In: B. Hilton (eds.) Emerging Spatial Information Systems and Applications, pp. 1–35. Idea Group Publishing, Hershey (2006)

Chapter 13
Accelerating Technology Adoption Through Community Endorsement

Richard E. Ullman and Yonsook Enloe

Abstract The idea that cross-discipline data interuse is a necessity for Earth systems science studies is now widely accepted, but data interoperability has been an elusive goal. The remote sensing community has a talented cohort of innovative developers of data and information technologies. However, despite the high level of innovation, progress on wide scale practical interuse has not been as rapid as expected. The usefulness of data interuse technologies requires adoption of these same visionary technologies on a wide scale. Thus, adoption of technologies to enable wider interdisciplinary investigation faces challenges similar to those for adoption of technology products in the commercial marketplace. Bringing a new technology from innovative use by visionary practitioners to popular use by more pragmatically oriented users has come to be called "crossing the chasm".

NASA's Earth Science Data Systems Working Group has been applying communications methodologies to bridge this adoption chasm within the agency. We have developed a process inspired by the successful model of the Internet "Request for Comments". The central idea of the RFC is notification to the wider community of specific detailed ideas that potentially affect interoperation of data and services though the Internet. Sharing ideas through the RFC mechanism has spurred adoption of Internet technologies and fostered collaboration in the development of Internet standards. NASA's Standards Process Group (SPG) seeks the same result in the domain of Earth science data systems.

Keywords Component · Best practices · Standards

R.E. Ullman (✉)
NASA/Goddard Space Flight Center, Greenbelt, MD, USA
e-mail: richard.e.ullman@nasa.gov

L. Di, H.K. Ramapriyan (eds.), *Standard-Based Data and Information Systems for Earth Observation*, Lecture Notes in Geoinformation and Cartography,
DOI 10.1007/978-3-540-88264-0_13, © Springer-Verlag Berlin Heidelberg 2010

13.1 Background

NASA's Earth Science Data Systems Working Group's (ES-DSWG) Standards Process Group (SPG) is one of several standing agency working groups charged with developing recommendations for the on-going evolution of NASA's Earth science computer data systems as a whole. The purpose of these working groups is to provide a way for data systems practitioners within the agency to provide input and feedback that will help guide the agency in the adoption and evolution of computer data systems technologies, software, practices and standards. The ES-DSWG's were formed in January 2004 (NASA ES-DSWG 2006).

13.2 How We Designed the Standards Process

After the deployment of the major components of the Earth Observing System Data Information System (EOSDIS), NASA looked to define the tenets to guide its continued evolution. EOSDIS is NASA's system of information systems for Earth observation data. It is multisite, featuring distributed data centers and comprehensive search, access, and distribution functions. EOSDIS was originally conceived as a monolithic system with the same data services at each installation, with the main distinction between the different data centers being the particular holdings and the discipline areas of expertise. However, even as EOSDIS was under development, it became clear that the different data centers had different data service requirements in the service of their different and particular communities. A NASA advisory committee was charged with identifying the principles for NASA's future data system. Under the working name "New Data Information Systems and Services (NewDISS)", the committee identified this diversity as a driving condition:

> "The [NASA Earth Science] research and applications community is extraordinarily diverse, with interests extending from the top of the atmosphere to the Earth's core. The standards and practices governing the acquisition, archiving, documentation, distribution, and analysis of Earth science are, de facto, those established by the disciplines specific scientific peer groups within this community. [The future EOSDIS] must recognize and embrace this tapestry of disciplines and subcommunities; there is no one-size-fits-all solution to the myriad data management needs of the community as a whole." [NewDISS Report, Page 13]

Following the NewDISS report, NASA instituted an activity to formulate implementation of this future EOSDIS. This activity, given the name "Strategic Evolution of Earth science Data Systems (SEEDS)" produced refined recommendations in topic areas identified by the NewDISS committee. The NewDISS committee recommended that standards for NASA's data systems be adopted by community consensus and that existing discipline specific practice drive the choices. The SEEDS recommendation termed it this way:

> "[The] data systems standards processes must enable participation by the community and by external organizations. Active participation in the [NASA's Earth science] data systems standards processes by the community, including data users, missions, value-added

> providers, application users, and data centers, is essential" [SEEDS Recommendation, Page 26]

The SEEDS study team looked at the practices of consensus-managed standards bodies to find a suitable model for a standards adoption process that would enable broad community participation with an emphasis on proving the effectiveness of existing practice.

> "For ongoing refinement, adoption and possible development of standards, recommend that [NASA] adopt a process similar to the Internet Engineering Task Force (IETF) process and tailored to meet specific [NASA Earth science data systems] needs. Develop a strategy for facilitating [Earth Science Enterprise] ESE standards compliance across the enterprise, including the performance of standards support services, e.g., user support, training, tool development. Encourage adoption of existing successful standards." [SEEDS Recommendation, Page 12]

From the IETF, the SEEDS study group took the central idea of a "Request for Comment (RFC)" broadly understood as a request to peers to consider the effectiveness of a particular practice. The IETF experience is attractive, because the initial explosive growth of the Internet was enabled by implementation of a set of relatively simple protocols by loosely coupled and diversely managed data systems. The NewDISS vision similarly prescribes a distributed management and loose coupling development paradigm for systems exchanging NASA's Earth science data. But the IETF RFC process has grown into a complex hierarchy of committees and the decision process is often long. The SEEDS study proposed that a NASA Earth Science Data Systems Standards Process Group (SPG) be formed to manage a RFC process tailored to match community nomination of standards with NASA's mission oriented needs. The SPG started work in 2004.

13.3 The SPG Standards Process

NASA data systems, especially those that support missions, must be reliable for high data volumes and for the particular data organization common to satellite-derived remote sensing data. And that means that managers of NASA data systems must be skeptical of technology usability experiences that may not reflect the challenges common to NASA Earth science data systems. In adapting the RFC idea, we have identified two levels of comment and review. First, we look for comments on the technology specification; that is, how successful has the community been in implementing the technology? Secondly, we ask stakeholders to comment on operational readiness. Does the technology promote interoperability or data interuse in an Earth science data systems environment? How well does the implementation of the proposed standard work in an operational setting? If the RFC is recommended as a NASA Earth science data systems standard by the SPG at the end of its process, it is because there is a defined community who has successfully documented the standard, implemented the standard, and has successfully demonstrated operational readiness with the standard.

The SPG considers recommendations for data systems practices. These are practices, technologies, or standards that increase the ability of NASA-generated data to be shared within and among communities of interest. Such practices include software application interfaces, data and metadata model conventions, data and information identification, common data services, formats, and other related technologies.

We use the term "communities of interest" or simply "communities" to include various stakeholders and affected constituents. Example communities include science discipline groups, users of similar applications, data systems developers, ESE mission planners, Earth science educators, data users, and others. Membership in each "community" often overlaps the others.

NASA's Earth science data systems standards process must facilitate interoperability between components of the Earth science network of data systems. Establishment of appropriate standards enables flexibility as future data and service providers have well-defined access points to join the network of data systems. To accomplish these goals, NASA's Earth science data systems Standards Process focuses on endorsement of practices that are relevant to Earth science network of data systems and that have mature implementations and proven operational benefit (NASA Standards Process Group 2006).

13.3.1 Organization

The Standards Process Group (SPG): This is the decision-making board of the standards process composed of part time permanent members from NASA's program office, Earth science mission projects, Earth science funded data systems awardees, and representatives from other agencies.

Technical Working Groups (TWG): These are groups commissioned by the SPG to conduct public review and evaluation of candidate standards, related implementations, and operational experience. Membership in a TWG is temporary and partially drawn from the Standards Process Group membership and can also be partly drawn from technical area experts and/or community members.

13.3.2 Path to RFC

The term "RFC" stands for "Request for Comment". The content of an RFC is either a technical note or a proposed standard. A technical note is any information that the submitter considers significant to the use of a practice within NASA's Earth science programs.

RFCs can come from any NASA stakeholder source including individuals that may be associated with or represent NASA's Earth science funded activities, industry or users of Earth science data. In some cases, the SPG may solicit an RFC. Other times, members of the community will bring forward an RFC to formalize NASA recognition or broaden use of standards that are used in their community.

The requirements for an RFC will be the same in each case. We require the RFC proposer to describe the practice or specification in technical detail or else provide references that describe the standard. The proposal must identify the community of use and citations of successful implementation and evidence of operational readiness must be provided.

13.3.3 Path to Community Endorsement

Figure 13.1 shows the steps from an RFC to endorsement as a NASA Standard. The process is characterized by technical analysis, open public review, and demonstration that the proposal "works". The first step is for the SPG to perform an initial screening and characterization. A TWG is assigned and a schedule is set, taking into consideration NASA need dates and support commitments. Also, any RFC must have two or more implementations before it can advance to draft status.

The TWG invites the community by means of email announcements to comment on the specification, operational readiness of implementations, and the usefulness of the technology and particularly to address questions formulated by the TWG. The TWG also identifies key stakeholders that are likely to have particular experience with the technology and solicits their opinion. The TWG reports to the SPG and the SPG makes recommendations for the final status of the RFC.

The role of the TWG is central to the review process. Because there is a wide variety of technologies that might contribute to interoperability or data interuse, the

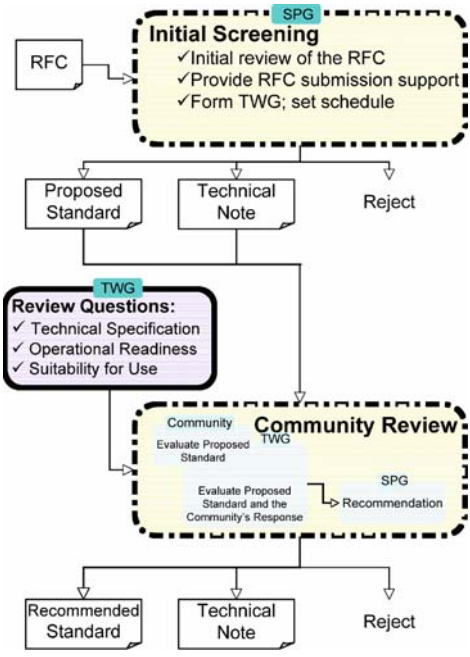

Fig. 13.1 Standards endorsement process

circumstances of each RFC are different. The TWG is the place where we wrestle with questions such as: What are the expectations for review of a specification? What evidence will show implementation? What does readiness for operational use mean for this RFC? And, what does suitability for use mean for this technology? The answers to these questions dictate the custom tailoring of our process to the particular RFC. The TWG forms particular questions to guide its evaluation of the RFC and to solicit reviews from and opinion of the community of practice. After some initial experience in approving proposed standards, the SPG has tailored its Standards Process to remove redundant reviews to shorten the review process. For example, if a proposed standard has already been approved by an international standards organization, the TWG may decide not to conduct a specification review. If a proposed standard is a *de facto* standard, widely used, and widely implemented, the TWG may decide not to conduct an operational readiness review.

The three types of reviews, the specification review, operational readiness review, and the suitability for use review, are conducted at the same time. The TWG decides if all three types of reviews will be performed for the candidate standard and may decide to perform a subset of the three reviews. For the specification review, the TWG asks reviewers to answer questions about accuracy of the specification in the RFC and to evaluate the significance of at least two implementations. The TWG must determine whether the implementations are independent and interoperable uses of the practice. For example, if the RFC proposes a format for a particular class of science product, demonstration of the use of that format by two separate "implementing organizations" would be considered two implementations. In the operational readiness evaluation, the TWG focuses on operational readiness of the implementations. Not only must the standard be demonstrated to work, but also the standard must be shown to work under conditions that are judged by the TWG and ultimately the SPG to assure that the implementations of the standard are robust enough to be ready for operational contexts of NASA data and NASA stakeholder users. The TWG also conducts a "usefulness for purpose" review, asking users to evaluate whether the proposed technology is suitable for a named purpose.

The TWG makes a recommendation to the SPG on whether to approve the proposed standard as a Recommended Standard by providing a written summary of the proposed standard along with the strengths, weaknesses, applicability, and limitations of the proposed standard. If the SPG approves the TWG recommendation, the SPG chair will send the written summary and recommendation to NASA HQ for endorsement of the SPG approved standard. The written recommendation of each endorsed standard is posted to the SPG website along with the RFC document. All future users of the endorsed standards have access to the documentation of the standard and a summary of the strengths, weaknesses, applicability, and limitations of the endorsed standard. The NASA Earth science community benefits from having this repository of endorsed Earth science data systems standards that have been successfully implemented and used within the NASA environment. The practical emphasis assures science investigators that standards proposed by members of a discipline community will directly contribute to science success in that specific discipline. We have seen that our standards process encourages consensus within a

community during the review phase. It can grow the use of common practices among related activities. But other discipline communities can use these same endorsed standards. Once a standard is endorsed, other discipline communities can learn from the successful practice and also use it. To the extent that documented use in one community can lead to adoption in others, this growth of common practices encourages cross-discipline interoperability. And the adoption of these common standards lowers the barriers to use of NASA data by external discipline communities within NASA and outside NASA and also lowers the barriers to entry into a network of NASA stakeholder data users and providers.

13.3.4 Standards Development

The ES-DSWG Standards Process is not a standards development process, but rather it is complementary to NASA's participation in consensus standards development organizations. NASA's involvement in organizations such as International Organization for Standardization (ISO), Federal Geographic Data Committee (FGDC), Open Geospatial Consortium (OGC), CCSDS, and other consensus standards organizations is designed to both contribute to technical excellence of the standards that are produced by these organizations and to assure the understanding of emerging technologies for early adoption within NASA. However, as we will discuss below, the development of new standards, even within the context of NASA representation that guides the technologies toward practical solutions is likely to fail to gain currency among NASA's mission planners. Some of the barriers are technical: the consensus, because it must satisfy the criteria of all participants, may be unsuitable for the particular demands of NASA's operational use. NASA data systems must be reliable for high data volumes and they must be applicable to the kind of environmental remote sensing and model data needed for NASA's science. These data tend to be global or regional in scope with moderate spatial and temporal resolution. But perhaps more important to the technical challenge is the challenge of convincing NASA stakeholder communities and particularly NASA mission and program planners that use of a given standard is worth the investment. The SPG Standards Process gives managers of NASA data systems assurance that any standard that is endorsed by the SPG will have high quality documentation as well as proven operational readiness. The SPG is likely to endorse only a small subset of the standards that ISO, FGDC, OGC, CCSDS, and other organizations develop.

13.4 "Crossing the Chasm"

The rollout and acceptance of new ideas or technologic products into a population or marketplace has come to be known in management theory as diffusion of innovation. The theories of this field focus of the sociological attributes of persons who choose to take up a new idea or product rather than the strength of the innovation itself.

Starting with a premise that in addition to the technological merit of the innovation, the sociological impact of an innovation is key to its success, the study of diffusion of innovation looks to the roles that different kinds of adopters play in the ultimate success or failure of the innovation.

For example, the acceptance of using a cell phone is less based on the quality of voice transmission of cell phones as compared with traditional phones and more based on the social benefits to a cell phone user to having non-tethered access to making and receiving phone calls. The market for cell phones first developed among customers for whom the potential of making a call on the road was overwhelmingly more important than voice clarity or even the higher reliability of wired lines. The cell phone represents a change in the culture of telephone use. Instead of being tied to a desk or looking for a phone booth, reliant on message-takers or subject to arbitrary hotel phone rates, the cell phone user becomes self-sufficient. Diffusion of innovation theory recognizes that to the adopter of a product that offers a sociological change, there is a particular risk, and that certain kinds of customers, because of their situations or goals are more willing to take such risks. In the case of the cell-phone user, self-sufficiency is not real; the cell phone user is subject to the service provider's rates and the completeness of the provider's coverage area.

The theory of diffusion of innovation is widely applied by marketers and analysts in the technology industry to explain why certain technology products are successful while others are not. This paper attempts to apply some of these ideas to the acceptance and use of standards within NASA's Earth science stakeholder community.

The commonly used model of innovation diffusion divides the population into five categories: Innovators, Early Adopters, Early Majority, Late Majority, and Laggards (Rogers 1995). While the exact measurement of each category as a percentage of the population may vary depending upon the particular innovation, the distribution plotting willingness to adopt a new technology against population is assumed to be similar to a Gaussian distribution bell curve (Fig. 13.2):

The categories of adopter are generally depicted as breaking along lines of standard deviation (Fig. 13.3):

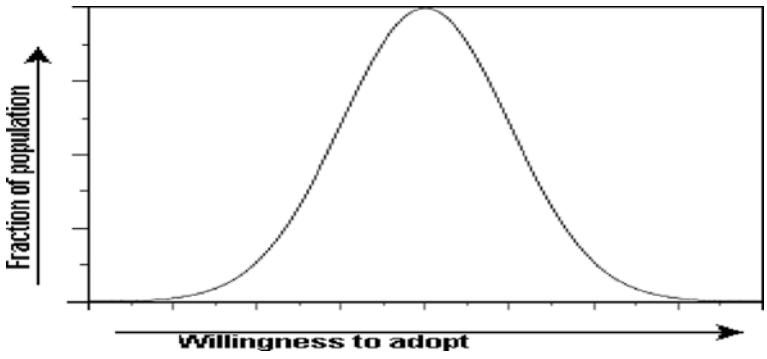

Fig. 13.2 Diffusion of innovation, propensity to adopt

Fig. 13.3 Categories of technology adopters

When depicted this way, it is clear that the largest part of the population, the mainstream, are either Early or Late Majorities, together these account for over two thirds of the total population. But by its nature, the progress of adoption of an innovation, the adoption must proceed from innovator to laggard over time. The cumulative acceptance or adoption of the innovation is notionally depicted as a cumulative distribution function (Fig. 13.4).

These graphs describe a tidy progression of an innovation from initial innovation to universal use. But, of course, this is only an ideal case. The cumulative adoption curve is the case for an innovation that successfully becomes diffused throughout a population. In fact, many innovations do fail to catch-on and when they fall by the wayside, the cumulative adoption over time abruptly stops. Even when considering a particular problem area or goal, multiple innovations might be tried to solve a particular problem. Most of those innovations fail to gain wide appeal. The result for the adopter of the failed innovation may be costly, as they must choose a different approach to achieving whatever goal to innovation was intended to accomplish (Fig. 13.5).

Geoffrey Moore (1999) proposes that one of the reasons for the failure of an innovation is poor marketing. And in particular, a failure on the part of organizations that are marketing an innovation to recognize the different decision motivations

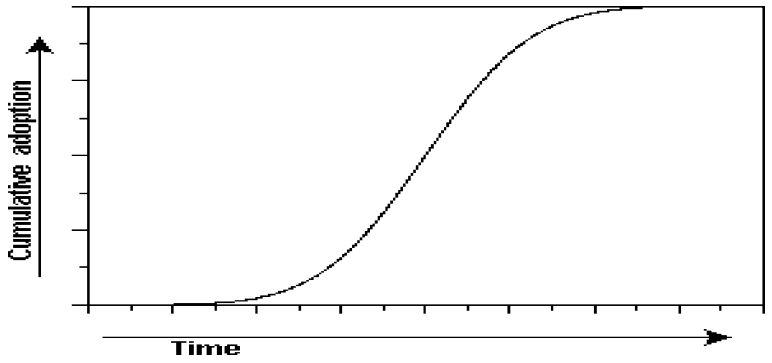

Fig. 13.4 Cumulative adoption over time

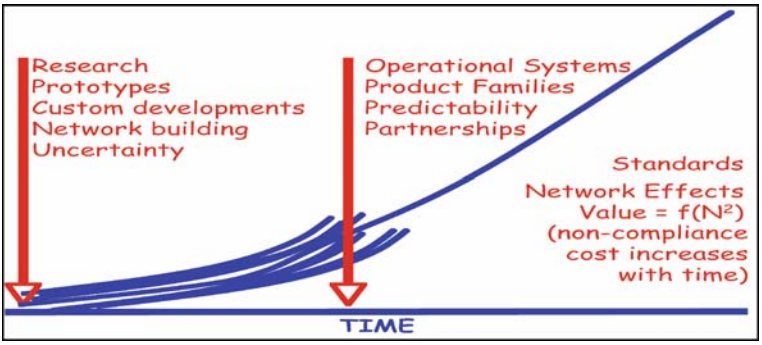

Fig. 13.5 Multiple innovations with final winner (Habermann 2006)

of the separate categories of innovation adaptors. In Moore's thesis, the different categories are in fact different markets. As defined by Moore, markets are:

- A set of actual or potential customers
- for a given set of products or services
- who have a common set of needs or wants, and
- who reference each other when making a buying decision.

The final two criteria give rise to gaps between the populations of different categories of adopter. And Moore's chasm is a profound gap between the Early Adopter and Early Majority populations' needs, wants and willingness to reference each other in buying decisions (Fig. 13.6).

Generally, the Early Adopters are looking to innovation to be a "change agent" that confers a significant new advantage or new opportunity. It is the pursuit of the new opportunity that is paramount, and Early Adopters recognize that the innovation is the key to achieving this new purpose. The Early Adopter is willing to put up with a certain amount of difficulty in implementing or operating the innovation because the new opportunity it enables is worthwhile. Also, the Early Adopter recognizes

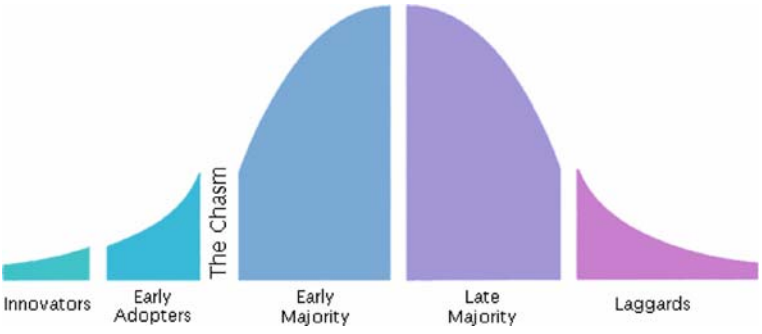

Fig. 13.6 Moore's revised technology diffusion model

that a certain amount of customization or extension is inevitable. Early Adopters are looking to press the innovation in order to achieve the new opportunities that they find latent in it. The adjective that describes the Early Adopter's relationship to an innovation is "visionary"

By contrast, the Early Majority is looking to adopt an innovation in order to achieve a productivity improvement. The Early Majority is not looking for a new kind of opportunity – they are comfortable with the success of the present system. They will adopt an innovation when there is a clear enhancement to current practice. Because of this focus on enhancement and productivity, the Early Majority will not be patient with bugs or extensive customization. As the "early" part of the majority, they understand that change and innovation are on balance both inevitable and good, but they want innovation to be as painless as practical and with known rewards. The adjective that best describes the Early Majority's attitude toward an innovation is "pragmatic".

So, using Moore's definition of a market, the Early Adopter and Early Majority are two separate markets because they differ in the third criteria; they have different needs or wants. What makes the gap between the two categories of adopters into a chasm is that they also have failed in the fourth criterion for a market. Their attitude toward innovation and their reasons for adopting an innovation are sufficiently different that the Early Majority does not reference the Early Adopter when making a buying decision and in fact, the Early Majority may distrust the recommendations of the Early Adopters.

The pragmatist knows that the visionary is willing to overlook or gloss over the flaws in a technology in order to pursue a vision. The pragmatist knows that the visionary is willing to disrupt the prevailing ways – In fact, that is exactly what the visionary proclaims. The pragmatist may find that the visionary is too enamored with technology that is not central to the mission or that the visionary fails to recognize the importance of experience or the importance of existing infrastructure. These are pragmatic concerns, and while a visionary might be able to demonstrate a successful use of a given innovation in one domain, the pragmatist is not interested in increasing uncertainty in his own project. There are plenty of challenges that the pragmatic project leader must face, and adding a potentially disruptive innovation makes the job harder and overall success riskier. For these reasons, the Early Adopter does not make a good reference for the Early Majority. The conundrum for those who wish to promote an innovation is that for a pragmatist, good references are a necessary precondition for adoption of an innovation. The largest part of Moore's book is dedicated to strategies to solve this conundrum.

13.4.1 The NASA Chasm

Moore points out that a promoter of an innovation can be lulled into believing that his innovation has crossed into the mainstream because of the adoption of the innovation by a visionary at an established institution looks outwardly like an adoption by a pragmatist. Large established institutions contain both visionaries and

pragmatists. In order to understand where an innovation exists in the diffusion life-cycle, one must examine the motivation for adoption. If the prime reason for use is as "change agent" then the adopter is an Early Adopter. If the reason for use is as a "productivity agent" then the adopter is a member of the Early Majority.

Indeed, NASA is no different. NASA projects have a range of technology adoption just as any population does. NASA's success often requires trail-blazing innovation. However, our mission areas have a pragmatic focus on reliability and safety that places projects and programs into a "majority" mind-set. So, within NASA Earth Science Data Systems, there are visionary projects that demand innovative technologies to achieve new purposes. There are also pragmatic systems that leverage existing infrastructure to continue to provide well-known services.

As a matter of course these separate markets within NASA's Earth Science broader domain will go about making decisions on what innovations to pursue and what ones to let fall by the way. But there are increasing pressures on to capitalize on Early Adopter innovation and press that innovation into mainstream service. These pressures are both visionary and pragmatic. On the visionary side, there are initiatives to qualitatively and quantitatively increase the sharing of NASA's Earth science data both within the agency and among other agencies and indeed internationally. In order to investigate the interrelated effects of various Earth science systems, data from multiple campaigns and missions must be combined in new ways. These visions cannot be achieved without the inclusion of systems designed and operated with pragmatic attention to existing agreements and investments. NASA also has pragmatic reasons to innovate. Standards have a well-documented effect to lower the cost of operation for complex systems. And NASA's Earth Science investigations are always cost constrained. There are always more desires for missions or investigations than can be funded. The greater the savings that can be realized by the use of standards in Earth Science data systems, the more fundamental Earth science research can be funded.

13.4.2 How to Cross the Chasm

Moore (1999) lays out a couple of tenets to guide the promotion of a technology product across the adoption chasm. Moore's prescriptions use marketing vocabulary from the perspective of marketing, with the promoter of an innovation selling to a market by influencing buying decisions. For NASA's SPG, we cast these same ideas in terms of standards proposals, communities, recommendations, and endorsement. Moore's solution in brief is:

1. Offer a "whole" product.
2. Become mainstream in a target segment of the market.
3. Branch out to related markets.

The "whole" product is defined as "the minimum set of products and services needed to fulfill the compelling reason to buy for the target customer". For a

pragmatist adopter, it is not enough that a core innovation is a technologically interesting component; the innovation must be embedded in a context of a set of products and services that do something practical. For an innovative technology, it is most often the case that the component does not drop neatly into an existing set of products. And so it is incumbent upon the proponent of the innovation to create the entire context, the "whole product".

The second point is that because pragmatists are looking for evidence that an innovation is practical, and the references they look to must come from other pragmatists, it is necessary to concentrate efforts in a niche of the larger potential market. An innovation must be a proven success, or else it is not of interest to pragmatists, and the best way to demonstrate success to a pragmatist is for others in their own reference market to have positive experience with the whole product. A promoter of a technology cannot afford to have an Early Adopter customer have a failure experience, and therefore, to assure success, the market must be small enough that it can be completely served. For reasons of definition of what constitutes a whole product, provision of technical and programmatic services and communication of references the initial market of an innovative product must be a consciously chosen niche.

Before an innovation can become eligible for the Early Majority market, it must become dominant in the niche market. This is so that there is overwhelming pragmatic endorsement of the product. It is not enough to have just a single success because part of the pragmatic calculation is that market leaders are safer choices. Only when there is market leadership in the niche, can use of an innovation safely broaden to the wider market.

In applying these same concepts to the SPG process and the endorsement of standards, the market is analogous to a community. The SPG has stated similar pragmatic principles (Ullman and Tsou 2005):

"Published practices that demonstrate benefit can grow . . .

- successful practice in specific community
- broader community adoption
- community-recognized 'standards'"

13.4.3 How the Standards Process Can Help Bridge the Chasm

Our goal in the Data Systems Working Groups, including the Standards Process Group, is to rapidly develop the "network effects" that characterize interoperability. The network effect is such that each additional use of the technology increases the value of all other users. A common prime example is the telephone, where each new telephone subscriber increases the utility of all telephones by permitting any user to access the new user. Also attendant with the increased unit usefulness is a decreased unit cost. As the components that participate in the network become more numerous, the technology for creating the component becomes better known and more easily replicated leading to lower cost for each additional unit. Data interuse

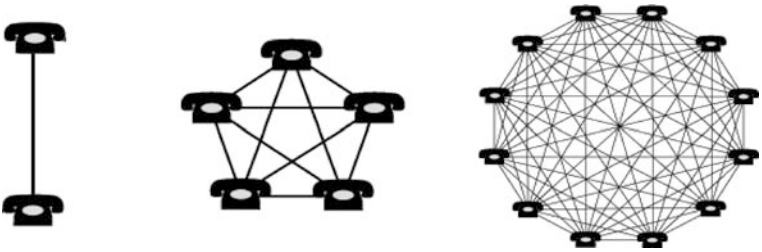

Fig. 13.7 Illustration of network effect (Coetzee 2009)

and data systems interoperability technologies should also exhibit these network effects. A primary goal of the Standards Process Group is to capture these cost savings and utility benefits by fostering standardized interoperability of both data systems and data products. To accelerate these network effects, we must accelerate the diffusion and adoption of innovative interoperability standards. And to do that we must understand the means by which such innovations become respected mainstream technologies (Fig. 13.7).

When we developed the Standards Process, we did not explicitly take into consideration the theories of diffusion of innovation nor did we develop the mechanisms of the RFC and community review as a "chasm-crossing" strategy for Earth science data systems standards. But in our experience thus far, we find that we are struggling with many of the challenges outlined in Moore's book. Moore's argument in *Crossing the Chasm* is that promoters of innovative technologies must understand that for the pragmatic Early Majority, references and relationships are of utmost importance. And the references must be pragmatic. Because, the visions of Early Adopters are not convincing to the pragmatist, there must be practical evidence that an innovative technology can do a whole job and can do it without unduly disrupting mission critical capability.

Moore further advises that successful chasm-crossing requires that the innovation win domination over a small specific market and use it as a springboard to adjacent markets.

Our standards process addresses these two aspects directly. While we have no financial motive if a particular technology "wins", for the network effect to provide increased value to NASA, *some* technology must dominate. The SPG process will force a pragmatic result. We seek to become, for NASA pragmatists, a reliable reference, and by seeking project experience from not only the direct innovators, but also the implementers and operators of a technology, we strengthen the word-of-mouth relationships that already diffuse technology within the agency. Secondly, our process emphasizes "community standards." We do not attempt to choose a particular technology, but rather we look to the specific communities within our broader group of Earth systems science stakeholders to relate their own experiences. If there is a technology that has begun to chasm-cross in that specific community, our RFC process is one of the springboards that can propel the practice to majority use among our stakeholders.

An endorsement process such as ours must always lag early adoption because we insist that before the SPG can recommend endorsement, the practice must have demonstrated readiness to work in NASA's Earth science operational context on a scale that is relevant for NASA's mission-focused Earth science data systems. Both observational data from NASA's Earth orbiting satellites and the output of model and simulation analysis are voluminous. NASA's EOS satellites generate several terabytes of new observations every day. Data transfers and analyses using these sources and assimilating other agencies similarly large data flows require highly reliable software and protocols.

But NASA Earth science data systems encompass multiple cultures of use. In research areas, there is a strong innovator and first adopter principle, while in missions there is a necessity for conservatism and reliance on proven operational principals. The Standards Process group endorsement process takes the valuable experience of the first adopter tendency of research activities and through careful, unbiased evaluation provides reliable guidance to data systems developers, pragmatists of the Early Majority who require dependable components for mission success.

13.5 The OPeNDAP RFC – A Community Practice Chasm-Crossing Example

In summer 2004, the Open Source Project for a Network Data Access Protocol (OPeNDAP) Group submitted the Data Access Protocol (DAP) v2.0 specification as a candidate community standard. DAP is a data transmission protocol designed specifically for requesting and transporting science data across the web using a client/server framework. On the server side, it organizes data from various common and custom file formats and storage models into name-datatype-value tuples. On the client side, the protocol relies on the widely used and stable HTTP and MIME standards. DAP provides protocols to accommodate gridded data, relational data, and time series, as well as allowing users to define their own data types.

The DAP protocol framework is particularly suited to simple access to remote data for scientist users. For serving data, not only are there freely available servers implementations designed to access local data in common formats such as HDF, netCDF and others, but also the API's in common programming languages C++, Java and Python allow data providers to build custom servers to accommodate specific dataset needs. For clients, data of any local type are presented using the common DAP data model and client API. For the scientist user there are clients of various sorts. DAP can be accessed using common web browsers, but for scientist end users, DAP clients within analysis applications provide a familiar and efficient method of accessing remote data stores.

DAP was originally developed as a component of "DODS, the Distributed Oceanographic Data System." A need for a data sharing and interoperability infrastructure was identified at a workshop of physical oceanographers at the University of Rhode Island in 1992. Using NASA and NOAA funding, an architecture was

developed, and the initial DODS system was developed over the next few years. Importantly, DODS objective was not just to facilitate access to data whether held at the local scientist's facility or in remote archives but crucially to allow a user of the data to analyze data using the application package with which he or she is the most familiar. These application packages are diverse, from general-purpose data manipulation software to unique discipline specific applications. And the oceanography community uses data from a variety of sources and in a variety of formats. The solution to the oceanography data sharing and interoperability problems was divided into two parts. One is discipline specific: the particular set of oceanographic data providers that together form a "virtual data system", a loose network to provide services to the oceanographic community through specialized clients, coordination on data formats, and particular datasets. The other is a discipline-independent core infrastructure for moving data over the Internet. By 1992, the World Wide Web had been proven to be robust and DODS data access infrastructure was built on the Internet's core standards. It was not long before the DODS developers realized that the discipline specific data services and the discipline neutral data access protocol infrastructure required different skills and resources to manage. The OPeNDAP, a not-for-profit corporation was formed to manage the development and maintenance of software implementing the DAP. The establishment of OPeNDAP freed the DAP protocol that was already technologically discipline neutral, from the specific discipline community that had conceived it. Scientists whose work crosses disciplines began to apply and promote the benefits of access to remote data stores from within familiar analysis application in other Earth science disciplines.

By 2004, when the NASA SPG began to look for community led standards, the DAP had experienced ten years of increasing use in sharing data not only for oceanography, but also other fields. DAP 2.0, a stable, second generation capability, was well established with years of operation of servers within the initial virtual oceanography data system network and among climate, weather, and environmental modeling centers and with scientists around the world, using both general and specialized clients within commercial and open source analysis applications including Matlab, IDL, Ferret, GrADS and others. OPeNDAP, representing this community of use, responded to the SPG's invitation to write an RFC with precise specification of the protocol, and enthusiastic members of the user community actively canvassed for input to the standards review process. The DAP specification (ESE-RFC-004) and comments received by the Technical Working Group may be found at http://www.esdswg.org/spg/rfc/ese-rfc-004

The SPG performed a standards endorsement process of the DAP 2 specification and decided to recommend it as a NASA endorsed Earth science standard. The review found particular strengths, both in the enthusiastic endorsement and significant operational use of the standard and in the clarity of the written technical specification and evidence of high quality and cross platform/cross data format implementations. Despite the endorsement, the SPG noted some less-significant technical weaknesses that might be addressed in future expansion or development of DAP. These include some difficulties in developing new DAP clients or in modifying components, in particular DAP implementations. There are some parts

of the DAP that some potential users felt could be improved. These shortcomings sometimes resulted in decisions not to use DAP for particular purposes. All requests for enhancements that were expressed in the SPG's review were forwarded to the OPeNDAP. The SPG found that DAP provided significant opportunities to expand availability of NASA data, with a large number of existing data systems and initiatives. As might be expected from the history of DAP, these systems are concentrated in the oceanography and atmospheric science communities. The SPG also identified opportunities to use DAP to maintain interoperability with other interoperability technologies that are currently in development or use by Earth science communities. These include most notably computation and storage grid technology components and OGC technologies. The SPG did recognize a "marketing" issue associated with DAP. DAP is closely linked in many users' minds with NetCDF, a data format common in the communities that first adopted DAP. This perception derives from the historical association. SPG considered a risk that, even though the specification is data format neutral, providers that use formats other than netCDF will be reluctant to implement DAP interoperability. The hope of the SPG is that endorsement by NASA would encourage these providers to participate. There is evidence that this SPG endorsement has had this benefit.

We believe that the SPG review has fostered the increased interest in using DAP 2.0 by the mission-success elements within NASA. While DAP had already begun to cross from one discipline community to another, the results of the SPG review encouraged managers of NASA's Goddard Earth Science Data and Information Service Center and NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) Data Processing System (MODAPS) and others to install DAP servers. We have also noticed an increased success of responses to NASA funding opportunities by those that propose to use DAP as part of their data management. From these anecdotal results, we believe that the SPG endorsement of DAP 2.0 by highlighting the pragmatic early majority use of the technology has contributed to lowering the barriers to the acceptance of DAP 2. In reviewing the experience, the OPeNDAP Group was asked about the benefit of the SPG process for the DAP 2.0. "The OPeNDAP Board of Directors singled this activity out as one of the most important for the past year. They felt that the benefits were well worth the (low) costs" (Gallagher 2005).

In hindsight, the endorsement of the OPeNDAP Group's DAP v2.0 specification by the SPG is a good example of the principles for success proposed by Moore in "Crossing the Chasm". Whether by design or not, the DAP followed Moore's prescription very closely. The new technology, an innovative use of web standards, was conceived by technology enthusiasts and conceived to address a particular need in a particular market. The market at first was the early adopters, technology enthusiasts and visionaries within the oceanography research community who were willing to put up with the tribulations of installing a new and untested technology either because they enjoyed the technology challenge itself, or because they saw in it a way to strategically achieve their business goal, that is, more effective research in oceanography. The technology was integrated into a complete solution, a "whole product" that addressed the needs of the broader market of oceanography research

scientists. In fact, these scientists were able to do their work more efficiently while avoiding the cost of operating outside the familiar environment of their favorite analysis tools. These scientists were not the visionary early adopters, but were the practical-minded early majority within their organizations. These pragmatic scientists were won over by the focus with the DODS developers on providing this whole product and because the product was developed with focus on their particular style of work. The DAP developers worked very closely with the physical oceanographic community to "win" the oceanography market first before branching out to other fields. After the DAP standard was firmly established and accepted within the physical oceanographic community with a cadre of practical endorsements the DAP standard and its associated software began to gain broader use within the Earth science community The physical oceanographic community became the pragmatic reference that technology worked in practice.

The OPeNDAP Group recognized that the SPG process has potential to facilitate even wider adoption of the DAP standard. With encouragement from the OPeNDAP satisfied DAP customers provided numerous and excellent quality technical and operational reviews. The Implementation Review and the Operational Readiness Review demonstrated to the wider community that this technology does indeed work and work well. Having a base community, like the physical oceanographic community, that provided proof that the technology works was invaluable to the adoption of this technology across a wider community. The OPeNDAP experience clearly illustrates Moore's argument in *Crossing the Chasm.* The DAP standard won the product domination over a small specific market, the physical oceanography community. The OPeNDAP Group used the domination in the physical oceanography community as a springboard to adjacent markets.

13.6 The WMS RFC – A Consensus Standard Chasm-Crossing Example

After the success in reviewing DAPS as a community nominated standard, the SPG next considered the OGC Web Map Service (WMS) specification as a proposed standard. On the surface, the RFC process appears similar. A representative of the OGC proposed that NASA endorse WMS based on the strength of the OGC-led consensus standards development process and the multiple implementations demonstrated through the OGC test-bed vetting process. Many organizations, including NASA and a large number of NASA stakeholders participate in OGC and the active participation of these organizations in both developing and demonstrating the standard lends credence to the proposition that the WMS is a viable community consensus standard. The OGC representative served a function similar to the OPeNDAP's representative with respect to DAP. It was not long into the review process that the SPG found significant differences in the way that the community responded.

The WMS specifies a protocol for server and client interaction to request and receive "maps" over the Internet. A "map" is defined by the specification as "a

visual representation of geodata; a map is not the data itself" (de La Beaujardière 2002). Three operations are defined: GetCapabilities returns a description of the information available from the server and particular request parameters; GetMap returns the geographically bounded map image; GetFeatureInfo returns tags that provide information about features shown on the image. The protocol runs over web protocols using URL syntax and responses are sent using XML. If a client requests different map images using the same projection and with overlapping boundaries, these separate maps can be "layered" on top of each other. The use of transparent background allows the layers beneath to be visible. The specification allows for a network of servers, each providing different layers to be combined into unique layered views by a client.

NASA was actively involved in the development of WMS from its earliest conception. NASA's collaboration with OGC began when NASA initiated a cooperative agreement with the organization in 1994, then called the OpenGIS Consortium. NASA's funding allowed the organization to grow and to develop a robust technical governance process. The goal of the cooperative agreement was to model interoperable visual environments and demonstrate the OpenGIS principles. The terms of the cooperative agreement include NASA direct in-kind participation as well as funding. Though the focus of NASA participation changed over the years and the office maintaining the collaboration shifted, NASA continued to fund and to participate in OGC testbed activities and technical steering. In 1997, an OGC paper, "WWW Mapping Framework" (Doyle 1997) provided the first concept that followed through the OGC process to a testbed activity name and eventually to the specification itself. NASA participated in every aspect of this process, funding participation in the testbed and providing the editor and a significant number of the technical staff who developed the specification document. NASA's involvement came out of its own applications division that is charged with enabling decision support applications of "national priority" by using appropriate NASA data assets for more effective outcomes. And the testbed scenarios and the specification's examples illustrate the use of satellite data, such as atmospheric imagery layered over land-covered imagery and terrain combined with political and infrastructure maps in such applications. From this intense NASA involvement, and the validation of scenarios developed and endorsed as NASA-relevant the OGC and the SPG logically concluded that NASA was institutionally committed to the success of WMS and that the WMS was operationally suitable for NASA data.

Curiously, the review process for WMS was not as smooth as the DAP review had been. While the SPG did find considerable general awareness of WMS within NASA and NASA's stakeholder community, few enthusiastic supporters emerged. The SPG began with an RFC from a user of WMS 1.3, describing the benefits of WMS for NASA data interuse and seeking comment on NASA stakeholder experience. However, during the Implementation Review phase, we discovered that most of the NASA and stakeholder projects had experience with the WMS 1.1.1 version and not the WMS 1.3 version. As described above, our process requires that the community as represented by review respondents identify positive experience with at least two instances of a practice, and so the SPG could not endorse WMS

version 1.3. The SPG instead used the path available to publish the RFC as a technical note indicating that it is provided as an example practice that is likely to contribute to interoperability or data interuse. This conclusion is justified by the technical merit of the 1.3 specification. After finding several instances of WMS version 1.1.1 installed by NASA and NASA stakeholder projects, the SPG issued a new RFC, requesting further experience with this version of the specification. The responses were a bit puzzling because several projects reported that WMS servers were installed and running, some even with many "hits" per day but that the capability was not "operational". By this, the stakeholders meant that the capability was not sufficient for the potential data volume, that the service was not provided for the full range of applicable data, or that their high-priority customers were not relying on the service. The SPG evaluated the use of WMS and the comments received and wrote an endorsement that balanced the acknowledged technical capabilities of WMS against the applicable uses and limitations. The major strength that SPG found with WMS 1.1.1 is that it is a mature specification with multiple proven implementations within many GIS application products. This is similar to the finding with respect to DAP wherein that protocol was embodied in familiar science analysis application products. WMS 1.1.1 implementations are embodied in GIS products including commercial market leaders such as ESRI and MapInfo as well as open source GIS such as MapServer and GeoServer. Many WMS servers exist, including those at US Government agencies, foreign government agencies, universities, and research organizations. Within NASA, many demonstration servers were identified. Stakeholders identified several weaknesses that the SPG found to be related not to the specification but to misapplication or misunderstanding of the technology.

NASA's overall experience with the Open Geospatial Consortium Web Map Service specification illustrates the difference between NASA's role in developing high quality standard specifications through active participation in consensus standards organizations, and the process leading to SPG recommendation as a NASA community standard. NASA's involvement in standards making can help to steer a standard to potential usefulness for NASA as an agency, but this involvement does not guarantee that such steering will be completely successful. In developing, especially a consensus standard, compromises are made. But even when a standard successfully responds to NASA's use scenarios, there is no guarantee that the NASA community will enthusiastically take it up.

Moore's theories of the adoption "chasm" might be instructive. Moore warns that when comparing the early adopter users of a technology and the early majority, it is on the surface difficult to tell the difference.

> "The reason the transition can go unnoticed is that with both groups the customer list and the size of the order can look the same. [. . .] But in fact, the basis for the sale – what has been promised, implicitly or explicitly, and what must be delivered – is radically different" [Moore, p 19]

The early adopter is the visionary. The basis for the sale, in our case the reason for participation or adoption of a particular interoperability practice, is the vision of a change agent that will solve an identified problem perhaps in a radically new way.

The visionary early adopter is willing to put in the work to deal with imperfections in the technology because they are committed to this vision. The early majority on the other hand is looking to evolve, not to revolutionize. What the pragmatic majority want most is the assurance of a productivity improvement that will not disrupt established practice. In the case of WMS, the NASA involvement in OGC represents a visionary participation. NASA's applications division is charged with finding innovative ways to apply NASA data to found applications. The mission data systems and data archive projects at NASA by contrast are pragmatists. Their first priority is to fulfill enumerated requirements carefully derived. In the case of serving maps, that is images of scientific data; any evolutionary improvement in efficiency is welcome, but such service is not a central concern. The participation in OGC by NASA early adopters did not logically mean that NASA early majorities automatically follow.

The history of the DAP was market domination in a single discipline community, what Moore would term a targeted segment of the market. In contrast, the WMS RFC did not first achieve market domination in a focused community, or at least not in a market in the sense of a set of NASA stakeholders who reference each other when making an implementation decision. Instead, the WMS came to the SPG with the only NASA stakeholder endorsement from early adopters. There were a few starts in the pragmatic market, but these did not rise to a level of operational endorsement. As far as the pragmatists were concerned, the sale had not been made. The WMS RFC was aimed at the broad market, broad within the scope of NASA data systems, of all map based geospatial data. Trying to capture the broad market all at once caused a very diffuse adoption of the WMS within the Earth science community. Thus, there was no identifiable community driving the efforts towards broader adoption of the WMS within the NASA Earth science community.

The result has been slower uptake of the WMS technology within the NASA Earth science community. But we do see that our SPG endorsement has had a pragmatic effect. With the demonstrated utility documented in our review of the standard and its endorsement by SPG as suitable as a standard, several NASA data centers (e.g. the Land Processes Data Center, the National Snow and Ice Data Center) are beginning to implement it within their data systems. It may be too early to tell if WMS has crossed the NASA chasm, competing technologies, especially geoTIFF in traditional GIS applications and Keyhole Markup Language (KML) coupled with Google Earth are competing for the application. The SPG's role is not to pick a winner; it is to serve as the pragmatic voice for standards adoption. And the SPG has not received RFCs proposing wider use of either geoTIFF or KML.

13.7 Impact

The ES-DSWG Standards Process recommendation is certainly not the only factor required for wider adoption of new data interuse or data systems interoperability technologies. But as NASA looks to rely more heavily on distributed systems under

distributed development fulfilling mission success, the widespread adoption of such standards is the only way to achieve the network effects necessary for cost-effective and flexible solutions. NASA is not lacking in innovative solutions, but successful adoption of those innovations will require crossing the same kind of chasm that faces marketers of new products. The Standards Process, by serving as a reliable reference and community of trusted sources can accelerate such adoption.

References

D. Coetzee, File:Network effect.png, http://en.wikipedia.org/wiki/File:Network_effect.png#filehistory, January 2006 [July 9, 2009].

J. de La Beaujardière, (Ed). "Web Map Service Implementation Specification" OGC Document #**01-068r3**, Wayland, MA:Open GIS Consortium, Inc January 2002.

A. Doyle, *WWW Mapping Framework*, OGC Document #97-009, Wayland, MA: Open GIS Consortium, Inc April 1997.

J. Gallagher, *Not Your Standard Deal: OPeNDAP's Positive Experience with NASA ESDS SPG*. Presented at the NASA ES-DSWG SPG meeting. San Diego. June 15 2005. Internet: http://spg.gsfc.nasa.gov/spgfolder/events/esdswg-meeting-june-15-2005/opendap.pdf [Apr. 20, 2006]

T. Habermann, *Innovation, Standards, and Mature Organizations*, Presented at the NASA ES-DSWG SPG Meeting, College Park, MD. November 15, 2006.

M. Maiden, et al. *NEWDISS A 6-To 10-Year Approach to Data Systems and Services for NASA's Earth Science Enterprise*, Washington, DC: NASA, 2002.

G. Milkowski, DODS: "Providing Direct Access to Distributed Research Data Resources." *Oceanography* (ISSN 1042-8275), Rockville, MD: The Oceanography Society, Vol. 8, No. 1, 1995

G.A. Moore. *Crossing the Chasm, Marketing and Selling High-Tech Products to Mainstream Customers* (revised edition). New York: Harper Collins Publishers, 1999.

NASA ES-DSWG. "Earth Science Data System Working Group." Internet: http://www.esdswg.org/spg Jan. 12, 2006 [Apr. 20, 2006].

NASA Standards Process Group. "NASA Earth Science Data Systems SPG – Portal." Internet: http://www.esdswg.org/spg Jan. 06, 2006 [Apr. 20, 2006].

OGC. "OGC History (abbreviated)." Internet: http://www.opengeospatial.org/ogc/history. Wayland, MA: The Open Geospatial Consortium, Inc. July 19, 2006. [Sep. 15, 2008]

E.M. Rogers. *Diffusion of Innovations* (4th edition). New York: The Free Press, 1995.

R. Ullman and M. Tsou. *NASA's Earth Science Data Systems Working Group Standards Endorsement Process, A Community of Practice Approach to Standards that Work*. Presented at the National Forum for Geosciences Information Technology, Washington DC, October 6, 2005.

S. Wharton, et al. *Strategic Evolution of Earth Science Enterprise Data Systems (SEEDS) Formulation Team Final Recommendations Report*. Greenbelt, MD: NASA Goddard Space Flight Center, 2003.