J Digit Imaging (2012) 25:213–226
DOI 10.1007/s10278-011-9428-4

Towards a Repository for Standardized Medical Image and Signal Case Data Annotated with Ground Truth

Thomas M. Deserno · Petra Welter · Alexander Horsch

Published online: 11 November 2011

Abstract Validation of medical signal and image processing systems requires quality-assured, representative and generally acknowledged databases accompanied by appropriate reference (ground truth) and clinical metadata, which are composed laboriously for each project and are not shared with the scientific community. In our vision, such data will be stored centrally in an open repository. We propose an architecture for a standardized case data and ground truth information repository supporting the evaluation and analysis of computer-aided diagnosis based on (a) the Reference Model for an Open Archival Information System (OAIS) provided by the NASA Consultative Committee for Space Data Systems (ISO 14721:2003), (b) the Dublin Core Metadata Initiative (DCMI) Element Set (ISO 15836:2009), (c) the Open Archive Initiative (OAI) Protocol for Metadata Harvesting, and (d) the Image Retrieval in Medical Applications (IRMA) framework. In our implementation, a portal bunches all of the functionalities that are needed for data submission and retrieval. The complete life cycle of the data (define, create, store, sustain, share, use, and improve) is managed. Sophisticated search tools make it easier to use the datasets, which may be merged from different providers. An integrated history record guarantees reproducibility. A standardized creation report is generated with a permanent digital object identifier. This creation report must be referenced by all of the data users. Peer-reviewed e-publishing of these reports will create a reputation for the data contributors and will form de-facto standards regarding image and signal datasets. Good practice guidelines for validation methodology complement the concept of the case repository. This procedure will increase the comparability of evaluation studies for medical signal and image processing methods and applications.

Keywords Signal processing · Image processing · Evaluation research · Data collection · Database management systems · Databases · Digital libraries · Electronic manuscripts · Image libraries · Medical Imaging Resource Center (MIRC) · Content-based image retrieval · Computer-aided diagnosis · Information system · Archive · Case repository · System architecture

List of Acronyms

AIM	Annotation and Image Markup
AIP	Archival Information Package
BIRN	Biomedical Informatics Research Network
CAD	Computer-Aided Diagnosis
CADe	Computer-Aided Detection
CADx	Computer-Aided Diagnostics
CARS	Computer-Assisted Radiology and Surgery
CAS	Computer-Aided Surgery
CBIIT	Center for Biomedical Informatics and Information Technology
CBIR	Content-Based Image Retrieval
CIP	Cancer Imaging Program
CLEF	Cross-Language Evaluation Forum
CMH	Cambridge Memorial Hospital

T. M. Deserno (✉) · P. Welter
Dept. of Medical Informatics, RWTH Aachen University,
Pauwelsstraße 30,
52074 Aachen, Germany
e-mail: deserno@ieee.org

P. Welter
e-mail: pwelter@mi.rwth-aachen.de

A. Horsch
Institut für Med. Statistik und Epidemiologie, TU München,
Ismaninger Str. 22,
81675 Munich, Germany
e-mail: alexander.horsch@tum.de

CMSP	Custom Medical Stock Photo
CT	Computed Tomography
DARE	Document Analysis Research Engine
DCES	Dublin Core Element Set
DCMI	Dublin Core Metadata Initiative
DDSM	Digital Database for Screening Mammography
DICOM	Digital Imaging and Communication in Medicine
DIP	Dissemination Information Package
DOAR	Directory of Open Access Repositories
DOI	Digital Object Identifier
DR	Digital Radiography
DRIVER	Digital Repository Infrastructure Vision for European Research
EFMI	European Federation of Medical Informatics
GIF	Graphics Interchange Format
GNU	Gnu is Not Unix
GPL	General Public License
GUI	Graphical User Interface
HUG	University Hospitals of Geneva
HTTP	Hypertext Transfer Protocol
IF	Impact Factor
IHE	Integrating the Healthcare Enterprise
IRMA	Image Retrieval in Medical Applications
ISI	Institute for Scientific Information
ISO	International Organization for Standardization
JPG	Joint Photographic Experts Group
LIDC	Lung Image Database Consortium
LONI	Laboratory Of Neuro Imaging
MEDIREC	Medical Image Reference Center
MRI	Magnetic Resonance Imaging
MIAS	Mammographic Image Analysis Society
MIE	Medical Informatics Europe
MIRC	Medical Imaging Resource Center
MOD	Magneto-Optical Disk
NASA	National Aeronautics and Space Administration
NBIA	National Biomedical Imaging Archive
NCI	National Cancer Institute
NIH	National Institutes of Health
OAI	Open Archive Initiative
OAIS	Open Archival Information System
OpenCV	Open-Source Computer Vision Library
PACS	Picture Archiving and Communication System
PDI	Preservation Description Information
PEIR	Pathology Education Instructional Resource
PET	Positron Emission Tomography
PMH	Protocol for Metadata Harvesting
PURL	Persistent Uniform Resource Locator
QBE	Query By Example
RepSIS	Repository for Standardized Medical Image and Signal case data references
RIDER	Reference Image Database to Evaluate therapy Response
ROI	Region Of Interest
RSNA	Radiological Society of North America
SciDR	Science Digital Repository
SDB	Simulated Brain Database
SIP	Submission Information Package
SSD	Solid State Drive
TCE	Teaching File and Clinical Trial Export
TCGA	The Cancer Genome Atlas
URL	Uniform Resource Locator
WGMIP	Working Group on Medical Image Processing
XML	Extensible Markup Language

Introduction

In a recent review on the current status and future directions of image processing in radiology that was published in the Journal of Digital Imaging, the authors conclude by providing strategies for moving forward in the next decade, which are the creation of standardized tools for data collection and the facilitation of data sharing [1]. Indeed, in signal- or image-based computer-aided diagnosis (CAD), there is also a strong need for medical reference data repositories [2–4].

Currently, medical signals and images routinely created in patient care are usually produced and archived digitally. Image processing systems have been developed for many uses, including computer-aided detection (CADe), diagnostics (CADx), and surgery (CAS). A few systems (e.g., computer-aided reading of mammography [5]) have been installed successfully in radiological practice, proving the high impact of medical image and signal processing on health care [6]. However, research progress that involves developing new algorithms in both fields advances slowly, and many ideas fail in clinical practice [7, 8].

To increase the acceptance of new methods, systems that process medical images and signal data must be successful in functional and performance testing as well as reliable in a comparison of algorithms. These criteria demand a solid and reliable evaluation based on "case data" that are preferably obtained from different modalities (e.g., computed tomography (CT), magnetic resonance imaging (MRI), digital radiography (DR), ultrasound imaging) and approved by medical experts. Here, the term "case data" refers to the combination of all of the images and signals in addition to the reference and clinical metadata that are required to evaluate a certain medical case. For example, in mammography, at least two views of each breast are acquired with DR and are sometimes complemented with MRI and/or ultrasound. For multimodal fusion, morphological (e.g., CT, MRI) and functional (e.g., positron emission tomography, PET)

volume data is obtained for each subject, sometimes at different points in time (e.g., pre- and post-operatively). This scenario often correlates with the term "study" in the Digital Imaging and Communications in Medicine (DICOM) protocol.

For the evaluation of image and signal analysis methods applied in the medical field, case data must be linked to a so-called gold standard or "ground truth". However, following a suggestion by Lopresti and Lamiroy [9], there is no such thing as "ground truth". Instead, there is the "interpretation" of data, which may differ naturally. There are no "right" or "wrong" answers, and hence, the ground truth is changed, modified, or corrected by data users, which aggravates the incomparability of the research [9]. In the following discussions, we continue referring to the "ground truth" when we address this type of information because it is a well-established term in medical image processing.

The collection of data is further impeded because of patient data privacy, copyrights, the availability of only small data sets without significant statistical validity, limited data selection facilities, a lack of reliable ground truth, and the high costs of reference data acquisition and pre-processing [10]. For example, a study on automatic bone age assessment is based on X-ray images, which are typically retrieved from a hospital information system. Approval of the ethics committee is required and access rights, anonymization, and transfer mechanisms, e.g., following the Integrating the Healthcare Enterprise (IHE) Teaching files and Clinical trial Export (TCE) technical framework, have to be established, which involve several organizations and departments [11]. Standardized metadata is established manually from the natural language text of diagnostic documents. This effort is immense but the resulting publications refer to the image processing algorithm rather than the data, which is usually not provided to other scientists. Aside from a lack of resources, this restriction results from the additional efforts of preparation of case data for public use. As a result, today's research is of limited traceability, which confines scientific progress. Novel algorithms and data processing methodologies are rarely compared with results published from other groups.

Nonetheless, medical case data repositories bear great potential for value-added services. Freely available case data can be disseminated widely and managed together with standardized metadata. System interoperability is maximized. Learning and teaching materials for physicians can be stored centrally to increase potential reuse and sharing. Usage metrics determine hit rates on data sets and document their usefulness. Altogether, standardized medical case data repositories provide cost savings and improvements in health care developments.

Numerous medical image and signal databases are available on the Internet [3]. A systematic search conducted in 2005 revealed that most of these databases at that time were commercial databases from the USA and UK (Table 1). The largest (Custom Medical Stock Photo, http://www.cmsp.com, Chicago, IL, USA) offers 1.2 million images from various medical domains including all modalities (e.g., CT, MRI, radiography, ultrasound, endoscopy, microscopy) as well as pictures from biology, pharmacology, botany, drugs, devices, and medical treatments. Registration and download are charged. The other databases were similar. None of them met the requirements for reference case databases as described above. Most critically, the ground truth was not provided, which is essential for validating and evaluating automatic image and signal processing.

The data repositories that were closest to meeting the requirements of a reference data source were significantly smaller (Table 2). Recently, a substantial amount of activity has taken place, indicating that the necessity of creating high-quality reference databases has now been widely recognized. Related initiatives that have emerged or have become increasingly relevant during recent years include the following:

- *NBIA*: The National Biomedical Imaging Archive (NBIA) at the Center for Biomedical Informatics and Information Technology (CBIIT), National Institutes of Health (NIH), USA provides access to a variety of complete image collections or ongoing projects for, e.g., CT Colonography, FDG-PET Lymphoma, Lung Cancer (the LIDC database, see Table 2), Reference Image Database to Evaluate Therapy Response (RIDER), or The Cancer Genome Atlas (TCGA) radiological images of genetically analyzed cases.
- *CIP Survey*: The 2010 Cancer Imaging Program (CIP) Survey of Biomedical Imaging Archives from the National Cancer Institute (NCI), NIH, USA created a survey of publicly available in vivo medical imaging archives and the underlying software capabilities, addressing lesion detection and classification, accelerated diagnostic imaging decision, and quantitative imaging assessment of drug response.
- *MIRC*: The Medical Imaging Resource Center (MIRC) platform (http://mirc.rsna.org) has been established by the Radiological Society of North America (RSNA), allowing for a cross-organization search of the so-called MIRC sites. Currently, the list of primary repositories ("Libraries") comprises more than 20 sites, including the Cambridge Memorial Hospital (CMH), University Hospital Geneva (HUG), Indiana University, MyPACS.net, National MIRC Singapore, and the University of Palermo. Libraries can be searched by text and/or patient attributes such as gender and year of birth. However, there is a charge for the download.

Table 1 The largest image databases in medicine ("Top 10") [3]

No	Name	Topic	Origin	Entries	Costs
1	Custom Medical Stock Photo (CMSP)	General	Part of mediaMD.com, USA	1,200,000	Image
2	Mediscan	General	Medical-On-Line Ltd, UK	1,000,000	Image
3	Medical Pathographic Library	General	The Wellcome Trust, UK	160.000	Image
4	Science Photo Library (SPL)	General	Michael Marten, UK	100,000	Image
5	Images.MD	General	Current Medicine of USA	50.000	Registr.
6	Pathology Education Instructional Resource (PEIR)	Pathology	University of Alabama, USA Dept. of Pathology	43,400	-none-
7	TICTAC	Pharma-cology	Virtual Health Network, UK	43,000	Registr.
8	Fleshandbones	General	Elsevier Ltd / Harcourt Inc, UK	30,000	Image
9	Medicalpicture	General	medicalpicutre GmbH, D	30,000	Image
10	MedPix	General	Uniformed Services University, Bethesda (USA)	19,997	-none-

- *Casimage*: The Case Image (Casimage) teaching files of the University Hospitals of Geneva (HUG), Switzerland (http://www.casimage.com) offer approximately 9,000 images from 2,000 cases. Images of various modalities with a strong focus on radiology have been collected and annotated with unstructured text in German, English, and French.
- *CLEF*: The imageCLEF campaign (http://www.clef-campaign.org) was started by the European Cross-Language Evaluation Forum (CLEF) initiative to evaluate image retrieval methods. Different datasets of more than 50,000 images were used. Typically, those datasets lack a proper ground truth.

Despite the common agreement on the need for standardized reference databases, resources are still rare, are often private, and are not accessible to other researchers. Those who have created the data prefer nondisclosure because distributing reference data does not yield an increase in scientific reputation, e.g., in terms of the Institute for Scientific Information (ISI) Impact Factor (IF). This fact, however, does not state the usefulness of scientific quality assessment by means of such measures.

In this paper, we aim at demonstrating that suitable and sufficient standards already exist, which allow standardized information repositories to be built that support the evaluation and analysis of medical signal and image processing algorithms and systems. First, we systematically analyze the requirements for a scientific case data repository. Based on the standards that were identified, we develop a system architecture that inherently embeds the required functionality. Based on its core modules, we show how such a system can be implemented successively, allowing rapid prototyping and immediate deployment. In the following, we refer to this approach as RepSIS—a Repository for Standardized Medical Image and Signal case data references.

Table 2 Medical image databases closest to the concepts of the EFMI initiative [3]

Name	Origin	Topic	Aim
Lung Image Database Consortium (LIDC)	USA	Lung cancer	Development and test of CAD algorithms
Digital Database for Screening Mammography (DDSM)	USA	Breast cancer	Development of CAD for early diagnoses
Mammographic Image Analysis Society (MIAS)	UK	Breast cancer	Development of CAD algorithms; test datasets and programs for comparing efficiency of algorithms
ERUDIT PapSmear Tutorial	DK	Cervical cancer	Early detection of cervical cancer; test and training dataset
Medical Image Reference Center (MEDIREC)	JP, G8	Cancer, cardio vascular diseases-	Reference images (DICOM, JPG, GIF)
Simulated Brain Database (SBD)	USA	Brain	Comparison of normal and diseased brain; 3D data for testing/comparing analyses
Biomedical Informatics Research Network (BIRN)	USA	Brain	Applications of medical image process-ing; three test environments
Laboratory of Neuro Imaging (LONI) Image Database	USA	Brain	Development of algorithms for mapping and exploration of brain functions

Materials and Methods

Towards Systemization

In 2002, the European Federation of Medical Informatics (EFMI) Working Group on Medical Image Processing (WG MIP) started an initiative to promote the establishment of a reference image database for medical image processing research and development groups, to support good validation and comparability of methods and systems (http://www.efmi-wg-mip.net/). There has been close contact with other initiatives, especially with the NIH, as well as industry (e.g., Siemens, Philips, Novartis) to develop the concept, which now embraces the following main components [3]:

- *Framework*: create an overall, economically sustainable framework for life cycles (define, create, store, sustain, share, use, and improve; Fig. 1) of reference case datasets and corresponding tools meeting the demands for validation and quality control of both academia and industry in research and approval processes.
- *Board of Experts:* establish a board of experts and enable them to define criteria for how to assess the relevance of a medical problem with respect to the importance of image and signal processing.
- *Key Applications:* using these criteria, assess medical problems and identify the most relevant problems that have a large impact on diagnostic and treatment outcomes.
- *Gold Standard:* specify case datasets that are needed for scientifically sound validation and evaluation for highly relevant problems, including quality criteria and standardization of data structures and metadata annotations.
- *Reference Data:* following these specifications, collect data and prepare validated case datasets to serve as common references for research and development.
- *Dissemination:* set up a platform for dissemination of reference case repositories, including bilateral cooperation agreements or contracts between provider and user with or without licensing, depending on the type of data and its use.
- *Quality management:* follow-up outcome indicators such as number and quality of published results, or costs and time reduction in the approval process, and the number of applications developed based on the references.

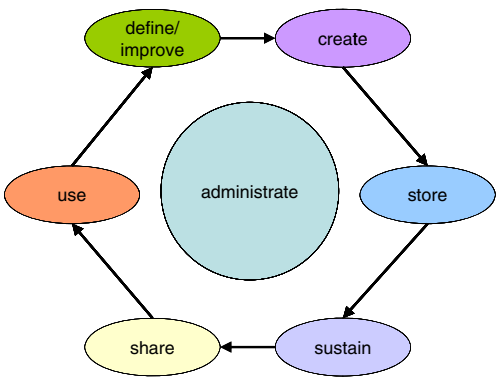

Fig. 1 Data Life Cycle. The colors refer to the OAIS model (Fig. 2)

During the past years, EFMI WG MIP has promoted these ideas, e.g., through several workshops and meetings at Medical Informatics Europe (MIE), Medinfo and Computer Assisted Radiology and Surgery (CARS) conferences [3, 8, 12]. The achievements comprise a template for provider-user agreements that includes clarifications about, e.g., the subject of the agreement, the user license, a binding to purpose, the time span, the reference character of the dataset, usage by a third party, media and copies, the reference copy, fees, co-operation, publication of the results, runtime, and irregular terminations [2]. Supported by these activities, datasets from projects in dermatology and gastroenterology have been made available to other researchers based on bilateral agreements.

Standards in Digital Archiving

To transfer such rather abstract requirements into a system's architecture, the use of standards is required. In the field of digital repositories, several standards and models have been established. The following standards are relevant and acknowledged:

- *OAIS Model (ISO 14721:2003):* The Consultative Committee for Space Data Systems (CCSDS), National Aeronautics and Space Administration (NASA) has defined a reference model for an Open Archival Information System (OAIS) that presents a framework for long-term repositories of arbitrary digital information and that has reached a high level of acceptance [13]. OAIS takes into account that a repository includes the provision of services and that the primary concern, aside from hard- and software, is in people, organizations, processes and work flows. The core concept is the information package, including the actual content and preservation description information (PDI), which support content understanding over the long term. The functional model embraces six entities:
  - *Ingest*: receipt of submission information package (SIP), pre-processing, and generation of archival information package (AIP);
  - *Archival Storage*: storage, maintenance and retrieval of AIPs;
  - *Data Management*: administering database, performing queries;

- *Administration*: overall operation, e.g., agreements with data producers;
- *Preservation planning*: ensure long-term accessibility of data;
- *Access*: support identification, control access, and deliver dissemination information package (DIP) to consumer.

- *DCES (ISO 15836:2009):* The Dublin Core Metadata Initiative (DCMI) has recommended the Dublin Core Element Set (DCES), a standard for the description of arbitrary objects containing 15 properties such as contributor, coverage and format [14].
- *OAI-PMH: The Open Archive Initiative (OAI)* has defined a Protocol for Metadata Harvesting (PMH). This protocol supports the exchange of metadata from repositories that is based on open standards such as the Hypertext Transfer Protocol (HTTP), XML and Dublin Core [15]. Harvesting comprises tools to gather, extract, index and search Internet information. It provides a means for automatic search and fosters the dissemination of data. OAI-PMH is supported by approximately 75% of the repositories worldwide [16].

Registries for digital repositories provide search engines and support discovery or the identification of digital repositories. Examples are the Directory of Open Access Repositories (DOAR), an authoritative directory of academic open-access repositories (http://www.opendoar.org) and the Digital Repository Infrastructure Vision for European Research (DRIVER) Networking European Scientific Repositories (http://www.driver-repository.eu), which has the vision to create a pan-European infrastructure for digital repositories.

Permanent digital object identifiers provide a continuity of references and ensure the unique retrieval of data over time. Examples are the Digital Object Identifier (DOI) and Persistent Uniform Resource Locator (PURL). DOI names specify identifiers for arbitrary data (e.g., text, audio, images, software). PURLs are web addresses that are not resolved directly but instead redirect the user to the actual location.

Standards in Medical Imaging

Several standards have been established in the domain of medical imaging and signal analysis, including the important issue of ground truth annotations:

- *ISO 21090:2011:* Health informatics – harmonized data types for information interchange (a) provide a set of data type definitions for representing and exchanging basic concepts that are commonly encountered in healthcare environments, and (b) specify a collection of 12 healthcare-related groups of data types suitable for use in many health-related information environments [17]. These standards are used by Health Level Seven (HL7) as well as Annotation and Image Markup (AIM) (see below).
- *DICOM SR:* Supplement 23 to DICOM specifies Structured Reporting (SR) classes for the transmission and storage of clinical documents. The SR classes fully support both conventional free-text reports and structured information, thus enhancing the precision, clarity, and value of clinical documentation [18].
- *AIM:* While DICOM contains large amounts of meta-data, it says little about the content or meaning of the pixel data. The ontology-based AIM initiative (https://cabig.nci.nih.gov/tools/AIM) is an emerging standard that makes the content of images accessible to computer applications that support automated processing. Within AIM, an "annotation" is regarded as explanatory or descriptive information about the pixel data that is generated by a human or machine observer, while an image markup consists of graphical symbols that are placed over the image to make an annotation. According to the core model, an annotation is built from an image annotation and an annotation of annotation, resulting in a structure that addresses hierarchically ordered annotations. Each image annotation may have one or more of the following geometric shapes: (a) point, (b) multiple point, (c) poly-line, (d) circle, and (e) ellipse [19]. The annotation is also linked to an imaging observation, which is defined from an imaging observation characteristic. This model allows standardized coding of ground truth and has recently been integrated in the clinical workflow for automated DICOM SR [20].

This short list emphasizes that both the concepts and implementations of image annotation models have recently become available and can now be adopted for use in a standardized case repository to define ground truth and to unambiguously link the ground truth to the pixel data for automated image processing.

Requirement Analysis

These preparations simplify the establishment of standardized medical image and signal reference case repositories. Case data from different medical fields that are sufficiently annotated with appropriate metadata and verified for use as ground truth shall be stored centrally and made freely available in a repository that guarantees long-term availability. In addition, RepSIS shall provide a considerable competitive advantage for data contributors and guarantee persistent availability of the case data, including total quality management.

Specifically, we aim at fulfilling all of the characteristics according to the final report of e-Science Digital Repositories (e-SciDR) [21]:

- *Selected*: the data held in RepSIS is limited to medical image and signal reference case data with corresponding

evaluations and proven ground truth. Other data will be declined.
- *Registered*: to make the case data discoverable, RepSIS will be registered in international repository directories (e.g., OpenDOAR, DRIVER). Also, we expect the RepSIS e-journal that holds the dataset descriptions according to the RepSIS requirements, which are authored by the data providers, to become frequently cited and recognized by reputation.
- *Managed*: RepSIS data is handled and stored in a defined manner, and a metadata description is supplied for proper administration. Data pre-processing that is inherent to the system will ensure that the RepSIS data conforms to managed quality criteria using standards such as OAI-PMH and DCMI.
- *Protected*: schemes regarding (a) access controls for authentication and authorization and (b) protection against intrusion will be established and implemented.
- *Sustained*: RepSIS data sets will be uniquely identified by DOIs. Any change in the data or the case set composition will automatically produce a new DOI. This rule will ensure and maintain the longevity, integrity, and authenticity of the data that is used in scientific studies.
- *Trusted*: the protection actions mentioned will strengthen the trust in the repository. This result will be enhanced by a follow-up on the repository's impact. All of the RepSIS data must be referenced in publications that personally credit (and/or institutionally credit) the data providers. This rule will increase the relevance, reputation, and recognition of RepSIS, making the repository interesting for data contributors. Trust will also be enhanced on the research side: the use of RepSIS will enable traceability and comparability, thus increasing the quality of the research.

Results

Based on a concise data model, we adapted the OAIS reference model. In the following, we first present the RepSIS data model, and then we present the RepSIS architecture.

RepSIS Data Model

- *Metadata Model*: Data producers inserting new case data must provide reference codes for their image and signal data as prerequisites before deposit. In addition, regions of interest (ROIs) and pathologies must be described according to the AIM standard. Data sets that are created must be described by metadata that is consistent with DCMI for harvesting from other repositories. Because case data can be selected from different providers, the metadata model must consider a comprehensive and unique definition of the data sets to guarantee reproducibility. This step includes all of the modifications on the images (e.g., scaling, color to gray conversion, and contrast standardization), which are logged in the metadata attached to the cases. This procedure aims at providing a unique data description.
- *Type of Dataset*: RepSIS aims at complying with the overall concept proposed by the EFMI, which distinguishes datasets related to four types of studies [2]:
  - *Simulator Study*: a dataset that is generated artificially by modality simulators, which are based on an algorithmic model of the modality's physical behavior.
  - *Phantom Study*: a dataset that is acquired from a well-defined, artificially prepared object with known measures such as size, position, density, and radiolucency.
  - *Elk Test*: a dataset that is regarded as a "minimum collection", comprising no more images or signals than are necessary to reflect the task-specific data variability with respect to modality, morphology, and pathology.
  - *System Trial*: a large volume of image or signal data that is suitable for a comprehensive validation of medical signal and image processing for CAD.
- *Data Definition*: Metadata management comprises insertion, storage and display of case data. Entries are locked for modification, to ensure uniqueness and integrity. The metadata includes the owner and the creator of the data set. If the creator has used data from the repository to build up a specific data collection, which is treated by the system as a new data set, then the information on the original owner is preserved. Appropriate interfaces enable the specification of metadata by the data provider. A history view provides insights that are useful in the creation of data sets and that are used for other purposes.
- *Ground Truth*: Appropriate upload mechanisms enable the specification of the image and the signal reference code. The approval of image and signal metadata by experts, to validate the data as ground truth, is enabled via specific interfaces and procedures. However, we model a one-to-many relationship from data to ground truth because we regard ground truth as an interpretation that naturally has differences [9], allowing ground truth alterations that are made by data users to be fed back into the repository. Hence, ground truth is seen as data, and RepSIS supports its entire life cycle as well.
- *Description Report*: Each data set is described by a standardized abstract that must be supplied by the creator or data provider. Requirements concerning the structure and content are defined by RepSIS and are

peer-reviewed by members of the board of experts. The abstracts are organized for browsing by users, thus supporting the selection of a data set. E-publishing in an appropriate portal for the publishing of research results online, such as an open access journal, is performed automatically after review.

RepSIS Architecture

The architecture of RepSIS is based on the OAIS reference model. Figure 2 presents the data flow. Data entities and processing steps are indicated with ovals and rectangles, respectively. The color indicates the modules according to the OAIS model.

The user enters the signal and image data that is annotated with metadata and the standardized ground truth description as well as the creation report for e-publishing as a RepSIS SIP. The entire submission process, including data upload, pre-processing, and binning, is managed by the Ingest entity.

The resulting RepSIS AIP is stored persistently. Specifically, the case and ground truth data and the history and change log data as well as the harvesting and metadata is managed by the Archival Storage entity. A permanent object identifier is created to uniquely identify the AIP. Repository registration and peer reviewed e-publishing of a data creation report, comprising a set of structured abstracts in the RepSIS e-journal, are produced automatically.

The Access entity manages the retrieval of suitable cases and the data download. Alternatively, the user can bin a data set from the case database, assisted by appropriate retrieval and search tools. Recommendations from best practices, assembled in RepSIS, support the data selection. Finally, the user receives a RepSIS DIP.

Other entities operate throughout the entire life cycle. Preservation Planning ensures the long-term accessibility of data. It controls the metadata of file formats, protocols, and hardware that is in use. The Administration entity is responsible for an efficient operation of RepSIS, which includes user support. The Data Management entity manages the databases and supplies the metadata that is used for harvesting by other repositories and for registrations in the directories of repositories. The Evaluation entity, which is added to the OAIS model, ensures a quality management cycle that results in the best practice guidelines. These guidelines affect the entire life cycle of the data, which includes the modification and binning of new data sets.

Ingest

According to the OAIS terminology, ingest refers to services that accept submitted data sets and prepare them for archival. It comprises (a) upload of data (SIP), (b) registration, and (c) pre-processing for quality assurance. The result (AIP) includes a metadata description and is ready for permanent storage. The procedure complies with standards including file formats and

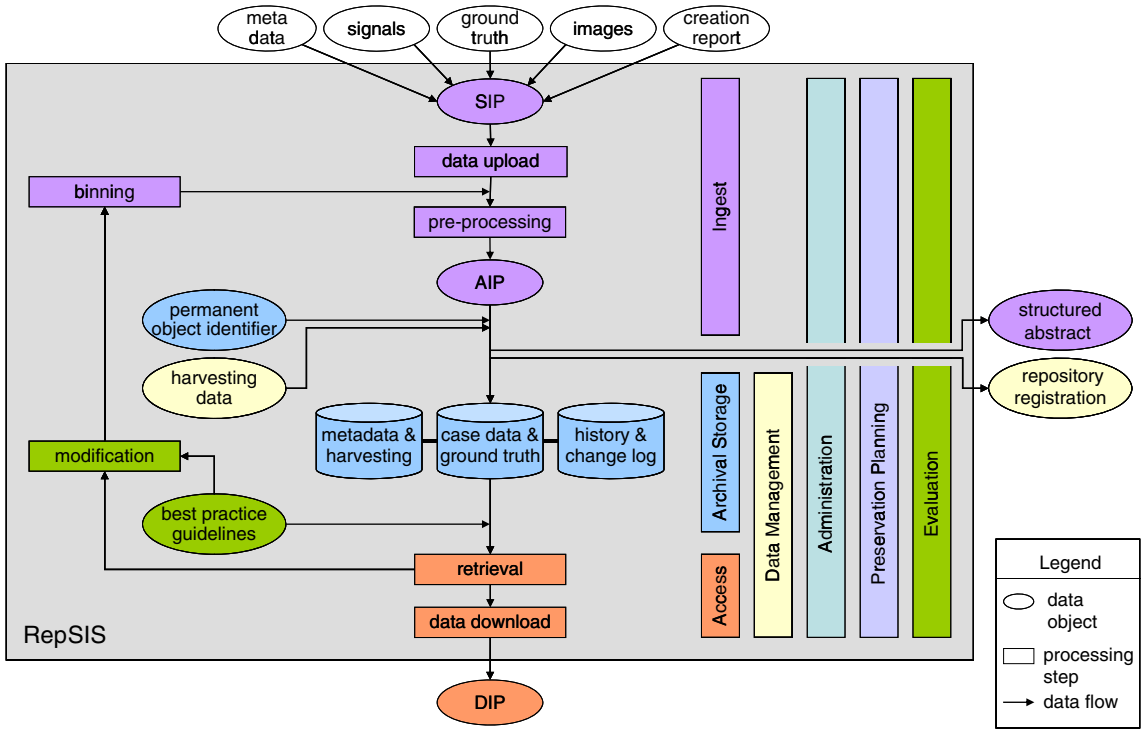

Fig. 2 RepSIS System Overview. The color indicates modules according to the OAIS model

data structures that are defined for the repository archive. The following modules are required:

- *Upload & Binning*: Data sets are established by the creation of new case data, ground truths, and the corresponding abstract, which includes a standardized description of the data content and creation. The RepSIS portal enables both data uploading and downloading. Novel data sets are also established by selecting and combining data from the repository. The definition of subsets and collections of data from different data sets is enabled. In the data model, user-specific creations are treated as new submissions but their case data is linked logically to avoid data multiplication.
- *Pre-processing*: The quality of the data is improved and harmonized by supplying standard methods for pre-processing before insertion such as checking a valid file format, converting to formats suitable for repository management and for repository users. Appropriate methods are adopted from open source libraries, e.g., the Open Source Computer Vision Library (OpenCV, http://opencv.willowgarage.com/) and MIRC services for anonymization. The repository portal provides an overview of internal methods and supports a convenient application.
- *Check & Approval*: The submission of new data and the definition of metadata have to be checked and approved by the board of experts before acceptance. Errors and inconsistencies would place the repository's data at risk. Actions such as data conversion or query management to request missing information are included in this module.

Archival Storage

This functional OAIS entity is responsible for the storage of the AIP into the appropriate area of the repository depending, for example, on the content of data. This entity makes entries in the appropriate database tables and puts files in appropriate directory structures. It handles the maintenance and retrieval of AIPs. The search index is updated, and error checking is performed. On request, AIPs are provided and delivered to the consumer. Furthermore, this entity supports maintenance and data retrieval, which includes the following methods:

- *Permanent Storage*: Based on the data model, permanent storage affects the methods that access the database tables and directories. Case data files are stored outside the database and have references pointing into specific directories. A search index is implemented that speeds up the search by separating it from the actual repository database.
- *History*: Modifications on the signal and image data by RepSIS internal methods, on metadata by the experts, or on ground truth by the data user are represented also in the metadata. Here, the mechanism of history logging is used [22].
- *Object Identifiers*: AIPs together with their creation report are uniquely specified by permanent identifiers (e.g., DOI, PURL). The management of the created identifiers will be implemented in a way that ensures high reliability.

Data Management

The primary functions of the entity Data Management include maintaining the database and its data, publishing to provide a means of locating repositories, and linking up with other repositories. The responsibility to address is to populate, maintain and access descriptive information about the repository's content, which requires the following modules:

- *Database Management*: The maintenance of the database is ensured by regular and thorough service checking, e.g., checking integrity and errors. A disaster recovery plan is established. After the implementation is completed, the performance will be observed and appropriate steps will be taken to increase the performance in case of decline.
- *Registration in Directories*: RepSIS will be inscribed in appropriate registries such as OpenDOAR or DRIVER, to link-up with other repositories. This accomplishment will increase the level of publicity for RepSIS. Harvesting by other repositories will be enabled using OAI-PMH. An appropriate concept is established that constitutes the attributes used in the harvesting of information. This concept is related to the metadata that describes the data sets.

Administration

The Administration entity manages the overall operation of the archival system. Its functions include the definitions of agreements. These are license models, policies, and terms of use [2]. The license model will be Open Access. This policy will grant free access and use of the repository data provided that the original authors and the repository are credited. Commercial use of the data is allowed, and in the case of modifications to the data, the use of the modified data remains under the Open Access license. This procedure complies with the Gnu is not Unix (GNU) General Public License (GPL, http://www.gnu.org/licenses/gpl.html) conditions for licensing software. Furthermore, the administration entity contains protection against intrusion. This entity is also responsible for the operation of the repository. The following modules are derived:

- *License Specification*: Aside from the specification of a legal valid license, policies will be defined for content, submission, data re-use, preservation, rights, and duties. The terms of use will define the conditions under which the repository will provide operations. According to Horsch et al., the agreement templates may include the following [2]:
  - Subject (license for the dataset, normally non-exclusive, perhaps time limited);
  - User license (the user is allowed to use the dataset in the framework of the scientific study referenced in this document);
  - Binding to purpose;
  - Validity time span (if applicable);
  - Reference character of the dataset;
  - Usage by a third party;
  - Media and copies;
  - Reference copy;
  - Fees (not relevant in our approach);
  - Cooperation;
  - Publication of the results;
  - Runtime and irregular termination of agreement.
- *Administration*: Operating a repository relies on the hardware platform and the technical infrastructure. The operation of the RepSIS prototype is administered. Specifically, a service plan is defined, and updates will be installed accordingly.
- *Documentation*: The overall operation requires thorough documentation of the technical and functional system. Special loggings and statistics are configured to provide sufficient insights on the reliability and security of RepSIS when operated.
- *Protection*: Data protection is assured by, for example, a firewall, anti-intrusion software, and data encryption. Appropriate actions are planned and performed that ensure protection of the repository data. During the operation of RepSIS, regular checks must be performed. Updates and new protection technologies are considered according to the protection plan that will be established.

Preservation Planning

This OAIS entity ensures that the stored information remains accessible over the long term. The environment of the repository is monitored and recommendations are provided. The tasks are mostly performed manually. Provider and users of the reference data are monitored, and adoptions that are required by the repository are planned. Technologies and standards are observed and migration to new development is designed (the database administration includes error checking and disaster recovery, as described in Data Management). The following modules are defined:

- *Monitoring*: Regularly, the employed hardware and infrastructure are checked as to whether they are still suitable for managing the repository's content, particularly with increasing use of the repository. New technologies will be evaluated.
- *Preservation*: Appropriate steps for preservation are based on monitoring results. This module also includes the implementation of adoptions that are made for preservation, for example, the migration of storage technology from magneto-optical disks (MODs) to semiconductor-based solid state drives (SSDs).

Access

The Access entity is responsible for the retrieval and extraction of data as well as for access control of authentication and authorization. This entity provides the dissemination information package (DIP) to the consumer. Downloads are registered for statistics on the usage of the data sets.

- *Account Management*: User management grants authorized access and monitors all of the downloads of case data sets. This task includes interfaces for user registration and modifications of account data.
- *Life Cycle*: RepSIS models the complete life cycle of data, i.e., define, create, store, sustain, share, use, and improve (Fig. 1). Every relevant processing step performed on the repository's data is logged, e.g., creation, modification, and download. In this module, the history view is modeled, providing details on the usage of the data during its life cycle in the repository in a compact manner, and also allowing comfortable searching.
- *Retrieval Support*: To discover the RepSIS case repository, the selection of data will be supported by sophisticated tools for content-based access, including convenient specification of search criteria and a comprehensive insight into the history of the data. The query and search tools follow the Query by Example (QBE) paradigm, i.e., retrieval by presenting a medical sample image.

Evaluation

During the operation of the repository, this unit reviews the success of the service and assesses the impact of RepSIS. The results are based on statistics from operation and log files. This entity also includes testing of the software and operation of the service. We supply a platform to exchange messages and information on best practices for the evaluation in medical research, including a user forum, statistics on case usages and dash-boarding of appropriate information. Approved

guidelines will be established and maintained. Specifically, the following modules are included:

- *Testing & Assessment*: This module contains testing of the implemented concepts, data structures and models. The assessment is a constant review of operations of the RepSIS prototype with increasing modular functionality. Different levels will be observed and suitable actions on improvements will be performed:
  - Technology-related level (performance, effectiveness);
  - Application-related level (ergonomics);
  - Content-related level (repository's concepts); and
  - Data-related level (best practice guidelines).
- *Operation Statistics*: The provision of log files will be a source for the assessment. Meaningful statistics will be set up from the log files and will be organized to support an automatic summary and interpretation.
- *Guidelines*: RepSIS provides a board of experts that monitor the repository and its use. Best practice guidelines will be released and will be quality managed to guide data providers as well as data users through the entire data life cycle: data creation and ground truth establishment, modifications, use, and report writing.

The Evaluation unit involves human experts who judge the progress and take appropriate measures for steering the entire life cycle of the case data. It is worth mentioning that this entity is not contained in the OAIS model but has been added to model the general life cycle of the data (Fig. 1).

RepSIS Implementation

In our prototype implementation, the resulting AIP is stored within the generic framework for content-based image retrieval (CBIR) in medical applications (IRMA, http://irma-project.org). The IRMA core components are (a) a database managing images and signals, algorithms and transforms, as well as computers and processing resources in the cluster [23], (b) a scheduler controlling distributed processing [24], (c) a web server providing a graphical user interface (GUI) [22], and (d) a communication interface interacting with medical information systems [25]. Specifically, the IRMA GUI logs the entire user interaction with the database and allows the restoration of any previous state [22].

The IRMA framework inherently provides algorithms and interfaces for CBIR data access that can be used directly to retrieve and compose novel data sets. A productive system, however, may be installed on a more reliable platform to guarantee data safety and security, accessibility and availability, and maintenance. Repository toolkits such as EPrints (http://www.eprints.org) and DSpace (http://www.dspace.org/about-dspace/introducing) are open source, OAI-compliant and approved in many repositories worldwide. The use of such toolkits provides an appropriate software platform including an adequate portal and will simplify the implementation. Typical repository functionalities and interfaces supplying basic operations are already given. The MIRC software offers tools for the exchange of data in clinical trials. On the basis of the selected tools, the cornerstones of the repository portal as a web interface will be laid.

Figure 3 shows an exemplary output of the RepSIS Feature Browser. The web-based GUI provides insight into the information on an image that is available in the repository. The medical image is shown in the top area, and the bottom area provides an overview of the corresponding ground truth and meta-information as well as information on all of the alterations that were performed over time. The example given in Figure 3 depicts a hand radiograph (Fig. 3a) that is used in three case data sets (IDs 969961, 903001, and 910863; Fig. 3b). The clinical metadata includes the patient's age (1.96 years), gender (male), and ethnicity (Asian) as well as the center coordinates of epiphyseal regions (Fig. 3c). Furthermore, five images showing the hand's single fingers, as well as the epiphyseal ROIs of each finger, are stored (Fig. 3d). More data on the pages that follow can be accessed by the right arrow sign in the navigation bar (Fig. 3e). This information includes the reading of two experienced radiologists and the results of commercial software for bone age assessment. These three interpretations set up the bone age "ground truths" of the hand radiograph against which algorithms may be benchmarked.

Hand radiographs typically are used for bone age assessment. The corresponding case data sets provide the developers of methods for bone age assessment with verified test data and a means for evaluation. The results of such methods can be reproduced by anybody. Furthermore, different algorithms can be compared to each other. The RepSIS case repository also allows a user to set up a new case data set, e.g., one containing only children younger than 2 years, to evaluate different algorithms concerning the bone age assessment of a very difficult age group.

Discussion and Conclusions

Enriching the OAIS model with the Evaluation entity in our RepSIS approach supports an unambiguous assignment of the data life-cycle with OAIS entities. This scenario is indicated by corresponding colors in Fig. 1 and 2. Furthermore, data creation is separated from the data definition, including the metadata, ground truths, and case data models, which comply with the ISO, AIM, and EMFI proposals, as well as from modifications that support data

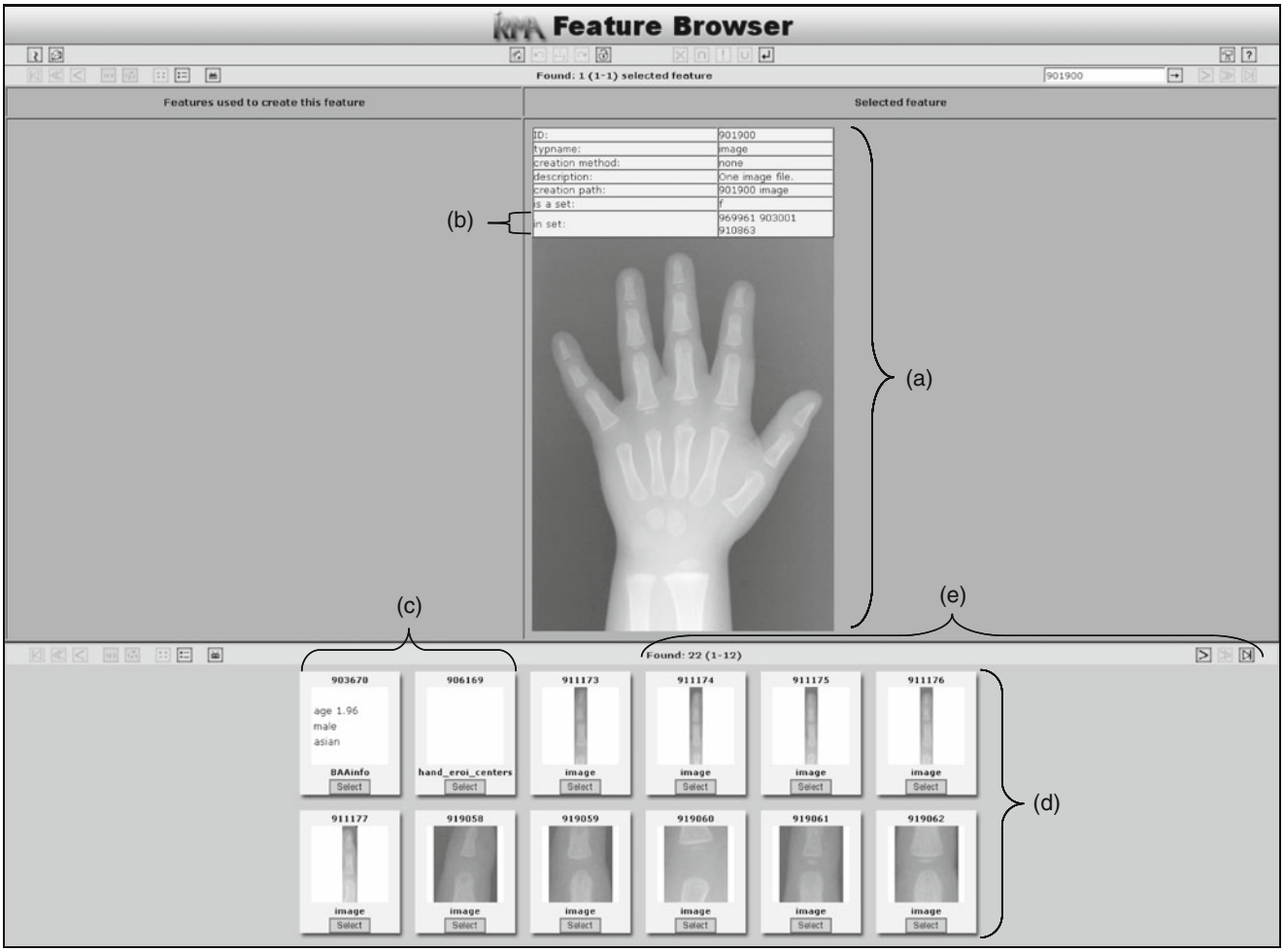

Fig. 3 Screenshot of the RepSIS Feature Browser. The annotations refer to (**a**) data element, (**b**) reference sets including the data element, (**c**) meta information associated with the data element, (**d**) transformed and modified versions of the data element, and (**e**) navigation to address further connected data

binning from different sources and providers. Herewith, the definitions and modifications are quality managed using the RepSIS guidelines.

Furthermore, a standardized data flow (Fig. 2) allows stepwise implementation and hence supports efficient engineering methods such as rapid prototyping. The following stages of implementation are considered reasonable:

1. The basic version encompasses a web presence of RepSIS, making available the foundation concepts and documents for the scientific community. Interested users and providers can connect through the web interface and will be informed about the goal and current status of the project.
2. The first productive repository version enables the insertion of case data and metadata as well as the creation and e-publishing of report abstracts. Modifications or compositions of data sets are not supported yet. Submission validation is performed manually. However, the download of data sets is already supported. Within the IRMA framework, this stage has been implemented already.
3. With this version, the repository becomes mature: it is registered in other repositories and exchanges data for harvesting. Permanent identifiers for data sets are introduced. Data sets from the IRMA database are migrated. Log files and statistics will supply a basis for assessment of the first phase of usage.
4. IRMA methods based on the OpenCV library are embedded for executing automatic quality checks of newly inserted data (e.g., ensuring the non-existence of doubles). Now, the ingest process is semi-automatic. Image pre-processing such as scaling and conversion to gray value is available. The history logs all data manipulations.
5. Datasets can be created as subsets from individual data sets and can be taken from different data sets with their own DOI. A forum for best practices on

the evaluation of medical image and signal data research has been set up.

6. RepSIS is transferred to a server platform, e.g., the SunSITE Central Europe (sunsite.rwth-aachen.de/) for continuous operation, and all concluding tasks (e.g., documentation, publication) have been performed.

Continuous extension of the repository, horizontally across modalities and medical problems, and vertically by increasing the size of datasets, is envisaged to take place in a productive, open research community. This organization encourages good practice benchmarking of new systems and methods in medical imaging and biosignal processing. With this concept, however, we did not yet address the resources that are required to maintain open data collaboration. Specifically, the development and support of infrastructure and the sustainment of the repository shall be assigned to an independent organization that is governed by public law, for example the National Library of Medicine (NLM) in the United States or the Institute of Medical Documentation and Information (DIMDI) in Germany.

The suggested RepSIS architecture is conformant to the typical three-tiered architecture of the user, with system (middleware) and the data layer, where the latter is based on the IRMA database engine. Data providers must transfer the data into the RepSIS archive, irrespective of whether the archive itself is modeled physically as a central or distributed system, and the RepSIS SIP structure must acknowledge the submission. Image pre-processing is offered within the repository. Recently, the Document Analysis Research Engine (DARE) project has proposed an open, decentralized, and community-driven model, allowing for storage, execution, annotation, and extension of data and algorithms without constraints. Within the cloud-based DARE architecture, data and/or algorithms remain local and are virtually connected by hosted application interfaces [26, 27]. While these paradigms seem contrary, they are in fact closely related. In the RepSIS model, data harmonization is performed on the conceptual level of RepSIS SIP, while DARE assumes that harmonization is completed on the protocol level of hosted applications. However, providing an interface to grant access to the users suggests that the data provider remains the "owner" of the data, but with the first access granted, a digital copy of the data has been created and provided "out of control". Furthermore, this concept bears the potential of inconsistencies; if a provider (temporarily) disconnects from the cloud, then data sets cannot be reproduced at any time.

However, the core DARE paradigm is based on several tenets [27] that in fact comply with the RepSIS concept:

- There is no absolute ground truth that is associated with the data, but instead, multiple and possibly contradictory interpretations can exist, which may vary with context and user. In RepSIS, we have modeled according to a one-to-many relationship for case data and its ground truth.
- All of the data and interpretations are persistent, fully queryable, and possess full provenance. The use of RepSIS DIPs for data dissimilation and addressing ReSIS data sets by means of PURLs ensure persistency.
- Data, interpretations, users, and algorithms have reputations on how they are accepted, rated, and used by the community. In RepSIS, the reputation is fostered by e-reports. Mechanisms to provide immediate feedback such as "like" and "don't like", which have been established already in the Web 2.0 of social networks (e.g., Twitter, Facebook), may add to RepSIS in the future, substituting the ISI IF of e-published creation reports.
- Automated evaluation results are third-party certified and evaluation conditions are reproducible. In RepSIS, we have included a board of experts as independent resources between the data users and providers.

In conclusion, we can state that the technology already exists to move toward the next decade of image processing in health care. Systems, architectures, and protocols are available to constitute reliable repositories of reference images and signals, including the enriched management of case data and reference data sets by means of CBIR [1, 28]. Comprehensive and reliable evaluation will yield more robust algorithms, and image- or signal-based CAD schemes will be incorporated into PACS and assembled as a package for the detection of lesions and for differential diagnosis [29, 30]. RepSIS is able to deliver an essential component to future work in medical imaging research.

Acknowledgments This research was partly funded by the German Research Foundation (DFG), grant no. Le 1108/9. The authors would like to thank George Thoma, National Library of Medicine, National Institutes of Health (NIH), USA, for his critical reflections on our approach and for contributing a perspective that improved this research.

References

1. Akgul C, Rubin D, Napel S, et al: Content-based image retrieval in radiology: current status and future directions. J Digit Imaging 24(2):208–222, 2011
2. Horsch A, Prinz M, Schneider S, Sipilä O, Spinnler K, Vallée JP, Verdonck-de-Leeuw I, Vogl R, Wittenberg T, Zahlmann G: Establishing an international reference image database for research and development in medical image processing. Method Inform Med 43(4):409–412, 2004
3. Horsch A, Blank R, Eigenmann D: The EFMI reference image database initiative: concept, state and related work. Procs CARS 1281:447–452, 2005
4. Horsch A, Hapfelmeier A, Elter M. Needs assessment for next generation computer-aided mammography reference image databases

and evaluation studies. Int J Comput Assist Radiol Surg 6(6): 749–767, 2011
5. Zheng B: Computer-aided diagnosis in mammography using content-based image retrieval approaches: current status and future perspectives. Algorithms 2:828–849, 2009
6. Andriole KP, Barish MA, Khorasani R: Advanced image processing in the clinical arena: issues to consider. J Am Coll Radiol 3(4):296–298, 2006
7. Dammann F: Image processing in radiology. Rofo 174(5):541–550, 2002
8. Horsch A, Punys V, Wismüller A, Castro Martinez A, Clarke L. Workshop on Validation of Medical Image Processing Systems at Tromsø Telemedicine Conference TTeC 2006. June 12, 2006, Memorandum, available at www.efmi-wg-mip.net
9. Lopresti D, Lamiroy B. Document analysis and research in the year 2021. Lect Notes Computer Sci 6703; 264–274, 2011
10. Müller H, Rosset A, Vallée JP, Terrier F, Geissbuhler A: A reference data set for the evaluation of medical image retrieval systems. Comput Med Imaging Graph 28(6):295–305, 2004
11. Kamauu AWC, Whipple JJ, DuVall SL, Siddiqui KM, Siegel EL, Avrin D: IHE teaching file and clinical trial export integration profile: functional examples. RadioGraphics 28:933–945, 2008
12. Horsch A, Thurmayr R: How to identify and assess tasks and challenges of medical image processing. Procs MIE 2003, 281–5
13. Consultative Committee for Space Data Systems (ed). Reference Model for an Open Archival Information System (OAIS). Blue Book CCSDS 650.0-B-1, Washington: NASA, 2002
14. Dublin Core (ed). Dublin Core Metadata Element Set, Version 1.1 2010-10-11, http://dublincore.org/documents/dces/
15. Lagoze C, van de Sompel H (ed). The Open Archives Initiative Protocol for Metadata Harvesting. Technical Report. Protocol Version 2.0 of 2002-06-14, Document Version 2008-12-07 T20:42:00, 2008. http://www.openarchives.org/OAI/openarchivesprotocol.html
16. Swan A, Awre C. Linking UK Repositories: Technical and Organisational Models to Support user-oriented Services Across Institutional and other Digital Repositories. Scoping Study Report, University of Hull, 2006, http://www.rsp.ac.uk/usage/harvesters.
17. ISO Technical Committee 215 (ed). Health informatics—Harmonized data types for information exchange. Technical Standard ISO/DIS 21090:2011.
18. Hussein R, Engelmann U, Schroeter A, Meinzer HP: DICOM structured reporting. Part 1. Overview and characteristics. RadioGraphics 24:891–896, 2004
19. Channin DS, Mongkolwat P, Kleper V, Sepukar K, Rubin DL: The caBIG annotation and image Markup project. J Digit Imaging 23(2):217–225, 2010
20. Zimmerman SL, Kim W, Boonn WW: Informatics in radiology: automated structured reporting of imaging findings using the AIM Standard and XML. Radiographics 31(3):881–887, 2011
21. The Digital Archiving Consultancy Limited (ed). Towards a European e-Infrastructure for e-Science Digital Repositories: a report for the European Commission. Technical Report 2006 S88-092641. Middlesex, UK, 2008. http://www.e-scidr.eu/wp-content/uploads/2007/03/e-SciDR_DAC_Final_Report.pdf
22. Deserno TM, Güld MO, Plodowski B, Spitzer K, Wein BB, Schubert H, Ney H, Seidl T: Extended query refinement for medical image retrieval. J Digit Imaging 21(3):280–289, 2008
23. Lehmann TM, Güld MO, Thies C, Fischer B, Spitzer K, Keysers D, Ney H, Kohnen M, Schubert H, Wein BB: Content-based image retrieval in medical applications. Methods Inf Med 43 (4):354–361, 2004
24. Güld MO, Thies C, Fischer B, Lehmann TM: A generic concept for the implementation of medical image retrieval systems. Int J Med Inform 76(2–3):252–259, 2007
25. Welter P, Riesmeier J, Fischer B, Grouls C, Kuhl C, Deserno TM: Bridging the integration gap from imaging to information systems: a uniform data concept for content-based image retrieval in computer-aided diagnosis. J Am Med Inform Assoc 18:506–510, 2011
26. Lamiroy B, Lopresti D. Supporting experimental research in computer vision. SPIE Newsroom 2001; April 1, doi: 10.1117/2.1201103.003558
27. Lamiroy B, Lopresti D, Korth H, Heflin J. How carefully designed open resource sharing can help and expand document analysis research. Proc SPIE 7874: 78740O, 2011
28. Deserno TM, Antani S, Long R: Ontology of gaps in content-based image retrieval. J Digit Imaging 22(2):202–215, 2009
29. Doi K: Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph 31(4–5):198–211, 2007
30. Müller H. Medical multimedia retrieval 2.0. Yearb Med Inform 55–63, 2008