# Information Technology Supported Convergence

George O. Strawn and William Sims Bainbridge

## Contents

Introduction ..... 280
The Rise of Information Technology and Its Impact on Science ..... 280
Data-Intensive Science ..... 282
Interoperability of Heterogeneous Data ..... 283
Creating "Knowledge Graphs" from Scientific Literature ..... 284
Knowledge Graphs: Science and Beyond ..... 286
A Range of Possibilities ..... 287
References ..... 290

**Abstract**

Modern information technology is transforming the collection, management, and sharing of scientific data in ways that greatly encourage convergence. Data-intensive science has evolved beyond the point at which all the information required for a research study can be centrally located, so interoperability across systems is required, with the additional benefit that data from different sources can be combined. Interoperability of heterogeneous data is a difficult challenge, requiring carefully specified metadata and well-conceptualized data management approaches like Digital Object Architecture. Scientific literature has become so complex and voluminous that it also must be managed in new ways, for example, using knowledge graphs to map connections as in Semantic

G.O. Strawn (✉)
Federal Networking and Information Technology Research and Development National Coordination Office, Arlington, VA, USA
e-mail: strawn@nitrd.gov

W.S. Bainbridge
Cyber-Human Systems Program, Division of Information and Intelligent Systems, National Science Foundation, Arlington, VA, USA
e-mail: wbainbri@nsf.gov



W.S. Bainbridge, M.C. Roco (eds.), *Handbook of Science and Technology Convergence*,
DOI 10.1007/978-3-319-07052-0_23

Medline. In the commercial realm, systems like Google Knowledge Graph and the related Knowledge Vault have begun to appear. For more than a decade, it has been recognized that future science will depend heavily upon distributed resources, including data archives, distant experimental facilities, and domain-specific research tools to enable new scientific discoveries and education across disciplines and geography. Similar approaches will become valuable for the development of abstract theory, for example, the cooperative construction of rigorous modular theories, in fields as diverse as physics and sociology, as scientists around the world contribute concepts and connect them by means of computer-based online tools.

## Introduction

The convergence of knowledge is the subject of this handbook. However, in science the convergence of knowledge is hampered by the *huge size* and *complexity* of science and of the scientific record. Regarding size of the scientific record, many scientists cannot even keep up with the new knowledge being created by their own field, let alone that of allied fields. For example, the US National Library of Medicine (NLM) maintains a database of the titles and abstracts of biomedical research articles called Medline (Kharazmi et al. 2014). This database currently contains 22 million articles. Regarding the complexity of science, there is disciplinary isolation created by independent technical vocabularies and non-interoperable data. This chapter will describe recent developments in information technology (IT) that are extending the scientific paradigm in ways that promise to increase interoperability and to support convergence by enabling connections among otherwise isolated knowledge fragments. Early realizations of these important developments will be highlighted.

## The Rise of Information Technology and Its Impact on Science

Modern IT might be said to have begun with the telegraph in the mid-nineteenth century or with punch card data processing at the turn of the twentieth century (Bainbridge 2004). However, the computer surely was the steam engine of the twentieth century. It emerged around the time of WWII and has increasingly influenced our technological society since then (Bainbridge 2006). The computer led to the Internet, which has integrated telecommunications into the digital revolution. In the early twenty-first century, the Internet has interconnected billions of people and computers as well as untold amounts of information.

Now the Internet is being augmented by the so-called Internet of Things (IoT), in which billions of currently connected people and computers will be joined by billions of connected devices such as scientific instruments, surveillance cameras, household appliances, and environmental monitors (Pew Research Center 2014).

These devices will contribute greatly to the flood of "big data," which science and society are already coping with. The IoT will accelerate the collection, processing, and communication of digital science data, but it could also increase the amount of non-interoperable data. Oceans of data will naturally connect the continents that represent major fields of science and engineering, thus promoting convergence. But convergence of multiple fields will also be required to create the IoT. For example, Jayakumar et al. (2014) have distinguished five types of IoT, depending upon their power supply needs:

1. Wearable devices such as smartwatches that must go several days between recharging.
2. Set-and-forget devices such as home security sensors that may operate for years without maintenance.
3. Semi-permanent devices such as sensors that monitor bridges and other public infrastructure.
4. Passively powered devices that lack batteries or permanent connections, such as smart cards carried in the user's pocket and activated by a reader machine.
5. Conventionally powered appliances, like smart kitchen microwaves, plugged into a power outlet while perhaps wirelessly connected to Internet.

A particularly promising new related trend is called *distributed manufacturing* from an industrial perspective or the *Maker Movement* from a popular perspective. The movement is a social phenomenon that resurrects traditional crafting hobbies through new technologies like 3-D printers, computer-assisted design, and online social media for sharing creative ideas (Axup et al. 2014). The Maker Movement is potentially far more than a hobbyist fad or an educational tool, as valuable as they can be, because it prototypes a form of manufacturing that could end reliance upon foreign industries and serve human needs better. In future, distributed manufacturing could create most products locally, customized for local cultures and conditions, in relatively small workshops employing local people who learned their skills in the Maker Movement, connected by information technology into the Internet of Things. Many fields of science and engineering must combine to make this vision practical, but absolutely central are computer-based systems for product design, coordination across a diversity of machines producing components from different materials, and management of the nationwide supply chain and local market.

Technology has always been an enabler of science. Early examples include Galileo and the telescope initiating modern astronomy and van Leeuwenhoek and the microscope initiating modern biology. Since the mid-twentieth century, IT has increasingly enabled all of science. The first electronic computers created "islands of computation," which quickly replaced armies of humans operating manual calculators. The most important science application that emerged at that time was the simulation and modeling of physical systems, which is now called computational science. Some observers have called it the *third paradigm* of science, placing it alongside theory and experimentation. Others have called it a new form of theory. Regardless of what it is called, most observers would agree it has had a profound effect on science. An even newer application of IT to science has been called *the fourth paradigm* by advocates (Hey et al. 2009). With less flair, it is also called

*data-intensive science* (Agarwal et al. 2011). The IT developments highlighted here are examples of data-intensive science and other data-intensive applications that promote convergence and connections.

## Data-Intensive Science

Data-intensive science emerged as computer storage capacities increased and costs decreased. First, it produced "islands of information" around large computers. With this development, huge output files, for example, from simulation runs, could be stored for later analysis. Then, as computers connected to the increasingly high-performance Internet, a "continent of information" was created. However, this continent consisted of heterogeneous data that were (and still are) largely non-interoperable. For the purposes of this chapter, *interoperable data* are those which can be employed together in computer applications. This is a problem in today's IT world because existing databases differ in almost every imaginable way, from having unrelated conceptual schemes for organizing the data to incompatible data storage structures, even just within one field, such as bioscience (Sansone et al. 2012).

Data can sometimes be made interoperable by time-consuming, expensive manual transformations. However, the goal of interoperability is to store scientific data in a form that such transformations can be performed automatically. One step in this direction is to add computer readable *metadata* – data about the data – to each data set. However, the form of the metadata must be sufficiently standardized to enable computer software to find and utilize it. Where interoperability has been achieved, such as with the Human Genome Project, major scientific advances have occurred.

The automatic interoperability of heterogeneous data will be realized when computers "understand" the data well enough to perform any required transformations. A similar understanding of textual information could aid science by greatly improving scientists' access to the articles most relevant to their research. Such an understanding could also advance science by enabling software to automatically deduce new results by combining results found in existing articles. Current computer-based keyword searches of huge databases have been a great step beyond manual searches. However, new IT developments are poised to take this activity as far beyond keyword searches as keyword searches are beyond manual searches.

A data-intensive society is also a defining characteristic in the early twenty-first century. As the World Wide Web was layered on top of the Internet in the early 1990s, the creation of Web pages exploded. Then search engines such as Google were developed to organize those pages to be able to respond to user queries. Those queries have traditionally been accomplished by means of keyword searches. Keywords are "meaningless strings of characters" (i.e., meaningless to computers), but they have been remarkably successful in locating pages that are often of use to persons performing queries. However, a second generation of search engines is emerging at this time. These new search engines can conduct "meaningful"

searches; that is, they focus on the *entities* referred to by keywords rather than the keywords themselves. Google's characterization of second-generation search is that the search is for "things, not strings."

## Interoperability of Heterogeneous Data

This section will focus on an approach to data management that lays a foundation for interoperability. It is called the *Digital Object Architecture* (DO Architecture) and was developed by Dr. Robert Kahn and his colleagues at the Corporation for National Research Initiatives (CNRI) in Reston, Virginia (Kahn and Wilensky 2006; Hoang and Yang 2013). Kahn was the codeveloper of the TCP/IP protocols, which are the foundation of the Internet, and the DO Architecture seeks to do for non-interoperable data what the Internet did for non-interoperable networks. Because of this analogy, a very brief (and partial) overview of the Internet architecture will be given.

The Internet is a *virtual network*, implemented only in software and riding on top of underlying "real" networks implemented in telecommunications hardware. The underlying networks are in general heterogeneous and non-interoperable. The Internet "stitches them together" with computers called *routers*, each of which is attached to two or more of the underlying networks. Ideally, every component and level of this set of networks needs to be optimized for data-intensive science (Dart et al. 2013).

One of the capabilities of the Internet is to enable the transfer of files from a computer on one network to a computer on another one. The World Wide Web, which has ridden on top of the Internet since the early 1990s, defined a protocol called Hypertext Transfer Protocol (HTTP) that enabled the convenient sharing of human readable information called Web pages (and now other applications such as e-commerce). In a sense, the Web provided for the human interoperability (i.e., information to be read by humans) of homogeneous data (Web pages which have a common format). A goal of the DO Architecture is to provide for *machine* interoperability of *heterogeneous* data. We now proceed to give an overview of the DO Architecture and to indicate how it can provide for such interoperability.

The Digital Object Interface Protocol (DOIP) is the DO Architecture analog to HTTP in the Web. The software of both of these systems can be visualized as an "hourglass," with application procedures in the top half and implementation procedures in the bottom half. At the narrow waist of the DO Architecture hourglass is DOIP, just as HTTP is at the waist of the Web. The Web defines both Web pages and URLs (identifiers) that resolve to Web pages, and the DO Architecture defines *digital objects* (DOs) and *handles* (identifiers) that resolve to digital objects. A Web URL is composed of a computer name followed by a "/" followed by a file name, and a handle is composed of a *prefix* followed by a "/" followed by a *suffix* (the prefix is assigned *to* an organization by the handle authority, and the suffix is assigned *by* that organization, but neither part is intended to be a semantically meaningful name). The DO resolved by the handle system differs from a Web page

in that it is the information *about* the data being referenced, not the data itself as it is in a Web page. In other words, the DO resolved by a handle contains *state information* about the data.

Two special types of digital objects are digital repositories and digital registries. All DOs are logically contained in a digital repository, and metadata for digital objects can be placed in separate digital registries or as part of a digital repository. When the handle of a DO is resolved, one of the pieces of information returned is the location of the digital repository that contains that DO.

As was stated above, the Internet can be viewed as a *virtual network*, implemented only in software and relying on underlying real networks implemented in telecommunications hardware. Similarly, the DO Architecture can be viewed as a *virtual database system*, implemented only in software and relying on underlying "real" database systems implemented in database hardware.

*Handles* identify data independently of the computer(s) where the data may currently reside. With proper management, there will be no "broken links" such as there are in the Web when a page is moved to a different computer. Another difference is that a DO is always *parsable*. That is, it can be "understood" by the software on an accessing computer because it is always in a standard form: a list of type-value pairs. Moreover, the types are also represented as handles and, therefore, can be resolved when the software does not understand them (but some software will be designed to expect and therefore to understand certain type elements).

HTTP laid a foundation for, but did not provide, many of the services that Web users expect today, such as easy-to-create and easy-to-read Web pages (via browsers), search engines to find relevant pages (Google, Bing, etc.), e-commerce sites (Amazon, United.com, etc.), and social media (Facebook, LinkedIn, etc.). Just as these applications have been developed in the upper half of the Web hourglass, many applications can be expected to emerge in the upper part of the DO Architecture hourglass. Because of this design, applications will have a built-in capability to establish interoperability of heterogeneous data. An example of the use of the DO Architecture to facilitate interoperability will be given below.

## Creating "Knowledge Graphs" from Scientific Literature

*Semantic Medline* is a *knowledge graph* created from Medline by Dr. Thomas Rindflesch and his colleagues at NLM. A knowledge graph can be defined as a *graph database*, which is a database in which the connections between the database elements are explicitly expressed (Pujara et al. 2013). That is, the elements are the nodes of the graph, and the relations between the elements are the arcs of the graph. For a convenient example in an overview of Semantic Medline, the team uses the example of *clock genes*, which as the name suggests manage time-related responses and are found apparently in all organisms, from fruit flies to humans (Rindflesch et al. 2011). In this case, the graph is a map of related concepts, which may belong to many different subfields of scientific research. Naturally, there are lines connecting "clock gene" with two specific genes, Cry1 and Cry2, which support

sensitivity to blue light and are involved in circadian rhythms that adjust behavior over the cycle of a day. But the graph also connects to some very human problems, including winter depression or mood disorders and even tumor growth. By connecting concepts, a graph such as this accomplishes an effective form of conceptual convergence.

A semantic graph database such as the Semantic Web also has built-in features to represent taxonomies and other hierarchical information structures. A graph can, among other ways, be represented in a computer as a collection of "triples" of the form (element, relation, element). Semantic Medline creates a knowledge graph from the text of the Medline abstracts. In each Medline abstract, there are "key sentences" which describe the results of the article. These key sentences can, in general, be restricted to the simple form subject-predicate-object. One of the contributions of Semantic Medline is a natural language processing module (NLP), which can find many of the key sentences in the abstracts. A related NLM development that is utilized by Semantic Medline is the Unified Medical Language System (UMLS), which is used to solve the *synonym problem* (Bodenreider 2004). That is, it provides for a *controlled vocabulary* which includes a unique identifier for each synonym class. These unique identifiers are used in the knowledge graph constructed from the key sentences. The controlled vocabulary enables the results of different articles to be put into a common language, thereby highlighting article commonalities. The subject and object nouns of each key sentence are nodes in the knowledge graph, and each predicate verb is an arc connecting its subject and object. On the average, three key sentences are discovered by Semantic Medline in each abstract, so 66 million key sentences currently constitute the knowledge graph.

The Semantic Medline knowledge graph can be both browsed and searched. Browsing via a graphical user interface enables an investigator to literally see connections among concepts and tie them back to the relevant abstracts. Graph search languages such as SPARQL enable scientific discovery by connecting isolated facts from the 22 million articles (DuCharme 2013). For example, the SPARQL query,

```
  Select "testosterone", ?relation1, ?x, ?relation2,
"sleep_problems"
  Where {
  "testosterone" ?relation1 ?x.
  ?x ?relation2 "sleep_problems".
  }
```

discovered two articles, the first asserting that testosterone *inhibits* the hormone cortisol and the second asserting that cortisol *causes* sleep problems. This discovery provided the first clue as to how decreasing testosterone in aging men might contribute to sleep problems. That is, no single researcher was aware of both articles.

As this ability to utilize the scientific record for science discovery becomes better understood, it will spread to other disciplines. One reason it has not been widely adopted so far is that the construction of language systems like UMLS is a labor-intensive process. As more automated methods for creating such language

systems emerge, this restriction will be alleviated. Assuming that the language systems across different disciplines can be properly articulated with one another, which will not be a simple undertaking given the international scope of science, interdisciplinary science discovery will be facilitated (Frade et al. 2012).

A second system for creating knowledge graphs from biomedical literature has emerged in Europe. Professor Barend Mons and his colleagues at the University of Leiden Medical Center in the Netherlands have developed *nanopublications* (Mons et al. 2011), in which the authors of publications identify and publish the key sentences as they write their abstracts. The requirement for a controlled vocabulary of concepts still exists, but no NLP module is required to find the key sentences. The Semantic Medline approach is very useful for the 22 million extant articles, but for new articles the nanopublication approach could be a viable option – if authors agree to take on the task of identifying their key sentences. Perhaps a hybrid approach will emerge, where Semantic Medline would be used to suggest key sentences to the author who could then accept or modify them.

## Knowledge Graphs: Science and Beyond

The "key sentences" described above as subject-predicate-object triples derived from text also have an interpretation as assertions about data, which enables knowledge graphs to naturally combine scientific literature and data into the same knowledge structure (Hebeler et al. 2009; cf. Cho and Kim 2015). This combination greatly enhances the possibilities of discovering new knowledge by mining the scientific record. As a simple example of how triples are used to describe data, consider a table of values where the rows represent experimental subjects and the columns represent specific attributes (e.g., one person per row and attributes such as weight and height in specified columns). In this context, a triple becomes subject-attribute-value, which is in the same "triple form" as key sentences. For example, a triple from such a table might be person1-weight-150, attributing a weight of 150 units to a specified individual human being. A controlled vocabulary representing the row and column names is required as it is for key sentences. Converting tables into triples is easier than converting text, and such conversions of "structured data" has begun to occur. For example, DBpedia has converted the tabular parts of Wikipedia into triple form (Bizer et al. 2009).

Any data table can also be represented as a DO. That is, a handle can be created that dereferences to a DO that contains state information about the table. That state information includes identifying the data as a table and indicating the number of rows and columns and the format of the data values. If the metadata also includes the controlled vocabulary information for the row and column headers, it would be a straightforward programming task to convert any such table into triples for a knowledge graph. At this time, however, the construction of the DO pointing to the data table may itself be a manual task. Assuming the DO Architecture comes into

general use, the creation of a DO for a data table could be automated, just as the creation of an HTML version of a document can be automated by a word processing system.

Finally, another possible use of the DO Architecture could be the use of handles to represent entities. As discussed above, UMLS and other entity systems determine the classes of synonyms (the entities) and assign a unique identifier to each one. In the Semantic Web implementation of a knowledge graph, the unique identifier is an International Resource Identifier (IRI). However, IRIs derive from Web URLs and hence at least appear to involve the names of computers rather than the computer-independent DO reference provided by handles.

As mentioned above, the effort to help computers better understand human intentions has moved into the quest for "second-generation search engines." This section focuses on Google's developments as an early example, but other vendors have signaled their intent to develop similar services (e.g., Microsoft announced that its Bing search engine will have a digital assistant called Satori Knowledge Base).

The Google Knowledge Graph and the related Knowledge Vault (Dong et al. 2014) will support three new dimensions for search: answer, converse, and anticipate. First, second-generation search will increasingly be able to answer questions rather than just identifying documents that may contain answers. (There currently exist several "answer engines," such as Wolfram Alpha, with similar goals.) Second, conversing might begin with disambiguation (e.g., "Do you mean jaguar the animal or jaguar the car?") and proceed to supporting additional search depth. Finally, Google can use the accumulated information from other searches to anticipate next search questions (e.g., "Previous searchers for jaguar the animal next searched for. . .").

As of 2012, the Google Knowledge Graph contained more than 500 million entities and billions of facts related to those entities. These numbers already dwarf the several million entities and 66 million facts that Semantic Medline has assembled. Thus, the "big data" dimension of computer semantics is being accommodated, just as the big data dimension Web search was accommodated by the development of novel "Web organizing" systems such as MapReduce developed by Google.

## A Range of Possibilities

A dozen years ago, the Interagency Working Group of Information Technology Research and Development identified a list of grand challenges, long-term scientific advances that require information technology breakthroughs and would benefit humanity. The first priority identified by this team was *Knowledge Environments for Science and Engine*ering, defined through these introductory sections of a substantial analysis (Strawn et al. 2003, p. 12):

Description of the Multi-Decade Grand Challenge
Organize and make broadly available distributed resources such as supercomputers, data archives, distant experimental facilities, and domain-specific research tools to enable new scientific discoveries and education across disciplines and geography
Focus in the Next Ten Years
Understand the needs of scientists and how science is changing (for example, data sets are more complex and teams are more interdisciplinary)
Increase access to computing systems, archives, instruments, and facilities
Build on successful experiments:
Upper Atmospheric Research Collaboratory (UARC) and Space Physics and Aeronomy Research Collaboratory (SPARC)
Network for Earthquake Engineering Simulations (NEES)
Biomedical Informatics Research Network (BIRN)
National Virtual Observatory (NVO)
Benefits
New discoveries across disciplines (for example, discoveries in one field can apply to other fields)
Establish new fields of science and engineering

Clearly, there has been tremendous progress since this report was published, and we have passed the end of the decade on which it primarily focused. Yet, this grand challenge could legitimately be made again, in essentially the same language, because progress has been a matter of degree, and we can reasonably imagine much more progress in the coming years. The discussion of this grand challenge went into some detail about what the technological challenges were, but its concluding section listed points that could be applied much more broadly (Strawn et al. 2003, p. 13):

**Indications of Progress**
More users of distributed science and engineering environments
More distributed science and engineering collaborations
More scientists and engineers in remote parts of the country
New tools and applications for more areas of science and engineering
New science and engineering ideas and innovations
Scientists and engineers achieve their goals more efficiently and effectively
More "hands-on" science education in K-12 and undergraduate school

The terms "distributed and remote" directly suggest convergence, and most "new tools and applications" would be valuable for multiple fields of research and development, thereby linking them. Convergence does not necessarily mean uniformity, however. This chapter has stressed the importance of connecting diverse sources of data that may have been assembled in different frameworks but have some commonality of conceptualization or domain. In many ways, human knowledge is far more chaotic that it could or should be, and much of the scientific and engineering effort needs to be invested in mastering that chaos. But an important part of success in that Herculean effort will be recognizing when fundamental commonalities do not exist and diversity must be maintained. For example, even within one field, there may be competing paradigms that would categorize data very differently. But that can be a good thing, because bringing the paradigms

together can result in new theory or theories in parallel rather than uniformity (Lewis and Grimes 1999).

A classic example relevant to the topic of this chapter is the constant but incomplete enthusiasm for propositional logic, production systems, or rule-based reasoning throughout the history of artificial intelligence. Alternative exists, such as neural networks of probabilistic methods. It is noteworthy that AI-pioneer Allen Newell (1990) titled his classic book on this topic, *Unified Theories of Cognition*, asserting that human and machine cognition could be explained by the same theories and directly promoting convergence. The general approach is to construct a system of propositions or if-then rules, based on clear definitions and axioms, logically deriving a potentially complex system of statements from rather simple elements. Superficially, this looks like divergence, but it actually achieves convergence of many empirically supported observations by finding a closed set of principles from which their complexity can be derived. Indeed, explanation becomes a form of convergence.

However, as in constructing a factory or a cathedral, much of the intellectual work of science as well as engineering requires assembly of many parts into larger structures. In a historically grounded analysis of theory in physics, Olivier Darrigol (2008, p. 196) describes this balance between divergence and convergence:

> any non-trivial theory has essential components, or *modules*, which are themselves theories with different domains of application. Even in alleged cases of reduction, modules remain indispensable because they play an essential role in the construction, verification, application, and comparison of theories. In this view, the heterogeneity of physical theory is best understood as modular structure; most of its unity rests on the sharing of modules.

Principles such as these can be applied to social science, as well as to physics and artificial intelligence. Barry Markovsky (2010) has applied the same logic to small group theory in sociology and social psychology, even citing Darrigol specifically. As George Homans (1950) pointed out in his classic *The Human Group*, individuals seek gratifying rewards but seldom can obtain them without help. Therefore, they form small groups of cooperating individuals, who come to conceptualize their aggregation as a valuable entity in its own right. Markovsky notes that modular theory requires development of explicit definitions of terms, distinct propositions stating meaningful principles, and logical rules for deriving hypotheses. Just as this approach can build a theory of small groups by assembling principles about individual behavior, principles about individuals and small groups can be assembled to produce rigorous theories of large societies. Within those vast social systems, hypotheses about science may be assembled with hypotheses about commercial institutions to produce a general theory of technological advance.

The fact that commercial search engine vendors are entering the field of computer semantics is a very good signal that the research phase of this field is about to give way to the early adopters phase, as modular theory might deduce. And these commercial systems will contribute to the development of additional systems for science like Semantic Medline, just as science systems will contribute to the further

development of commercial systems. In other words, another important public-private partnership is emerging in the dynamic IT industry. Similarly, the emergence of a commercial Internet of Things and the associated cyber-physical systems that are being built on top of it have a strong need for data interoperability, such as enabled by the Digital Object Architecture. Here, too, a public-private partnership is emerging that will serve the needs of both science and society.

The convergence of scientific knowledge is hampered by the size and complexity of science and the scientific record. We must improve our ability to find connections within and among the various science domains if convergence is to proceed unimpeded. As the scientific record, both data and articles, are digitized *and* made interoperable, an important barrier to the convergence of scientific knowledge will be reduced. Systems such as the Digital Object Architecture, Semantic Medline, and nanopublications will be applied to more, perhaps all, science fields, as well as fields beyond the science and technology enterprise, as the Google Knowledge Graph demonstrates. These systems, applications of the fourth paradigm of science, will increasingly contribute to the convergence of knowledge.

**Acknowledgments** This manuscript was written in conjunction with the NSF/World Technology Evaluation Center (WTEC) international study on Convergence of Knowledge, Technology, and Society. The content does not necessarily reflect the views of the National Science Foundation (NSF) or the US National Science and Technology Council's Subcommittee on Nanoscale Science, Engineering and Technology (NSET), which is the principal organizing body for the National Nanotechnology Initiative.

## References

Agarwal D et al (2011) Data-intensive science: the Terapixel and MODISAzure projects. Int J High Perform Comput Appl 25(3):304–316

Axup J, Thomas AM, Waldman A et al (2014) The world of making. Computer 47(12):24–40

Bainbridge WS (2004) Hollerith card. In: Bainbridge WS (ed) Berkshire encyclopedia of human computer interaction. Berkshire Publishing Group, Great Barrington, pp 326–328

Bainbridge WS (2006) Computer. In: McNeill WH et al (eds) Berkshire encyclopedia of world history. Berkshire Publishing Group, Great Barrington, pp 421–426

Bizer C et al (2009) DBpedia – a crystallization point for the web of data. J Web Semant 7(3):154–165

Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32:D267–D270

Cho SG, Kim SB (2015) Summarization of documents by finding key sentences based on social network analysis. In: Ali M et al (eds) Current approaches in applied artificial intelligence. Springer, pp 285–292

Darrigol O (2008) The modular structure of physical theories. Synthese 162:195–223

Dart E et al (2013) The Science DMZ: a network design pattern for data-intensive science. In: Proceedings of the international conference on high performance computing, networking, storage and analysis. ACM, New York, #85

Dong X et al (2014) Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 601–610

DuCharme B (2013) Learning SPARQL. O'Reilly Media, Sebastopol
Frade JR, Di Giacomo D, Goedertier S et al (2012) Building semantic interoperability through the federation of semantic asset repositories. In: Proceedings of the 8th international conference on semantic systems. ACM, New York, pp 185–188
Hebeler J, Fisher M et al (2009) Semantic web programming. Wiley, New York
Hey T, Tansley S, Tolle K (eds) (2009) The fourth paradigm: data-intensive scientific discovery. Microsoft Research, Redmond
Hoang T, Yang L (2013) Scalable and transparent approach to media archive using digital object architecture. In: Proceedings of the 46th annual simulation symposium. Society for Computer Simulation International, San Diego, Article 6
Homans GC (1950) The human group. Harcourt, Brace, New York
Jayakumar H et al (2014) Powering the internet of things. In: Proceedings of the 2014 international symposium on low power electronics. ACM, New York, pp 375–380
Kahn R, Wilensky R (2006) A framework for distributed digital object services. Int J Digit Libr 6(2):115–123
Kharazmi S, Karimi S, Scholer F, Clark A (2014) A study of querying behaviour of expert and non-expert users of biomedical search systems. In: Proceedings of the 2014 Australasian document computing symposium. ACM, New York
Lewis MW, Grimes AJ (1999) Metatriangulation: building theory from multiple paradigms. Acad Manage Rev 24(4):672–690
Markovsky B (2010) Modularizing small group theories in sociology. Small Group Res 41(6):664–687
Mons B et al (2011) The value of data. Nat Genet 43:281–283
Newell A (1990) Unified theories of cognition. Harvard University Press, Cambridge
Pew Research Center (2014) The internet of things will thrive by 2025. Pew Research Center, Washington, DC
Pujara J, Miao H, Getoor L, Cohen WW (2013) Ontology-aware partitioning for knowledge graph identification. In: Proceedings of the 2013 workshop on automated knowledge base construction. ACM, New York, pp 19–24
Rindflesch TC et al (2011) Semantic MEDLINE: an advanced information management application for biomedicine. Inf Serv Use 31:15–21
Sansone S-A et al (2012) Toward interoperable bioscience data. Nat Genet 44:121–126
Strawn GO, Howe SE, King FD (eds) (2003) Grand challenges: science, engineering and societal advances requiring networking and information technology development. National Coordination Office for Information Technology Research and Development, Arlington