Developing and Using Standards for Data and Information in Science and Technology

John Rumble, Jr.[1], Bonnie Carroll[1], Gail Hodge[1], and Laura Bartolo[2]

[1] Information International Associates, 1009 Commerce Park Drive, Suite 150, Oak Ridge TN 37830
jrumble@iiaweb.com
www.infointl.com
[2] College of Arts and Sciences, Kent State University, Kent, OH 44242-0001
lbartolo@kent.edu

Abstract. The Information Revolution has created the possibility of electronically preserving virtually all scientific and technical data and information indefinitely. The ultimate goal is to enable future researchers and other users to find, access, and use these data information for further scientific discovery and development of new products and services. The entire preservation effort becomes fruitless if data and information cannot be found, accessed or used. Considerable work has gone into the development of standards to achieve this. The standards address the actual data and information, the electronic formats in which they are stored, and the metadata used to enable finding and accessing the data. Significant barriers still exist with respect to the adoption and use of these standards. We explore several of these barriers, including those related to nomenclature, linguistic, sociological and economic, and technical issues. We conclude the paper with suggestions about ways that these barriers are being and in the future can be overcome.

1 Introduction

The explosion of information technology (IT) has led to its permeation into every corner of science and technology (S&T). The need for standards as a key technology to take advantage of the new IT capability is almost self-evident, but progress has been slow. Developing and using standards in a research and development environment is considerably different from an industrial environment. The need for standards may be just as great in both; yet the motivation and approaches have turned out to be significantly different for a number of reasons. With the advent and proliferation of mark-up languages, many data and information specialists anticipated that the process of developing S&T data and information (STDI) standards would become easier, and progress would be rapid. Indeed, in some sense, the formats of

these standards are easier to specify, but the content thereof, that is the semantics, of the standards remains as elusive as ever.

In this paper we examine how standards are developed and used in the area of S&T data and information. In particular we detail some of the reasons that progress has been slow and suggest mechanisms for overcoming barriers. We begin by examining the reasons standards are developed and their use in S&T data and information (STDI) work. We follow with a discussion of four types of barriers to progress: nomenclatural; linguistic; sociological and economic, and technical. We then discuss international aspects of STDI standards and conclude with suggestions on how some of these barriers can be overcome.

For the purposes of this paper, we define data as the qualitative and qualitative results of experiments, observations, theory and calculations, and information as all other ideas, interpretations, descriptions, expositions, and reports about experiments, observations, theory and calculations.

2 The Reason for Standards

Standards of all types are developed primarily for economic reasons. Once a specific solution to a problem has been found, or a specific method is found to be widely applicable, standardization offers considerable savings in terms of time, money and other resources. Internal to a particular organization, the promulgation of a standard can be relatively straight forward. When an organization attempts to impose a standard solution or approach external to itself, other interested parties must consent in some manner. The present-day industrial standard framework has arisen in response to the fact that most problems or processes have multiple solutions, and that to agree on a "standard" solution requires negotiation and agreement among parties with competing interests. When the economic benefits of a common (or "standard") solution are recognized, agreement is forthcoming, and standards are adopted.

Other reasons besides economics play some role in the development and use of standards. A standard approach or solution may be *intellectually superior* to what can be developed internally by an organization, as it describes an understanding and knowledge of state-of-the-art capability. Standards can also *codify and communicate knowledge* in a manner that can be used as a basis for further product or service development. A standard can clearly describe a *structure of knowledge, methodology, data or information* so that its application is easy and unambiguous, leading routinely to good solutions for recurring problems. Finally, a standard can provide an *accurate description* of a system, the methods used to study a system, including what was controlled and what was measured, that allows clear communication of a process, test or method.

All of these reasons come into play with respect to the development and use of S&T data and information standards. For example, standards for reporting data on measurements on protein crystallographic structures are of major importance in demonstrating all the above factors. [1] They reflect an *intellectually superior* approach to report the results of an experiment. They *codify and communicate* existing approaches in a manner that as new methods are developed, the changes and improvements can be clearly defined. [2] They provide a *structured method* for demonstrating the structure determination was performed correctly. They also define a clear *accurate description* of what was done in a manner that others will believe the determined structure.

Any standard for S&T data and information tries to accomplish all these goals, which explains why individuals and organizations are so motivated today to develop such standards.

3 The Use of Standards in S&T Data and Information

As shown in Table 1, standards are used for a variety of purposes throughout the life cycle of S&T data and information, from their initial generation through their long-term archiving.

Each of these uses requires emphasis on different aspects of the content and format of STDI standards. For example, for standards related to data generation, many of which are incorporated into automated data gathering instruments, the standard must cover a wide range of independent variables. As pointed out by Shoshani et al [3], in a given experiment, only some of these variables are varied. Others are set at the beginning of the experiment or observation and not varied. Rarely does an experimenter report all possible variables involved, so the standard must account for that fact. Because the use of these experimentally-generated data is often immediate, standards for the long-term storage of these data may not be necessary. That is not true for observationally-generated data as found in astronomy and the earth sciences, or for large scale physics experiments, in which data use can be done long after initial generation.

In contrast, standards for data archiving must include contextual information over and above the individually reported results. Archived data used years or decades later cannot be interpreted without a detailed explanation of the analysis used, the assumptions made in the experiment and other related information. The same holds true for all other stages of the data and information life cycles. The types and amount of metadata required vary, leading to varying requirements for the relevant STDI standards.

Table 1. Use of Standards for S&T Data and Information

Applications of STDI Standards	Uses
Data generation	Capturing results for experiments observations, and calculations; recording independent variables and context; developing protocols; assuring inter-experimental and – observational consistency
Database building	Database schema definition; database input; data uniformity; data consistency; database interoperability
Data evaluation	Data quality assessment; assessing completeness of reported data; comparison of different data sets; generation of reference data sets
Database use	Accuracy of retrieval; identification of needed data; data integration; repeatability of retrieval; accuracy in retrieved data
Data reporting	Uniformity; completeness in papers
Data access	Locating needed data; cross database use; search engine accuracy; abstracting and indexing; information location
Data archiving	Documentation of what and how the data was archived; support of migration of software and hardware; support of precision of recall
Data exploitation	Input into software; understanding of completeness; automated retrieval and use
Data visualization	Input into tools that allow visual analysis and inspection of data sets

4 Examples of S&T Data and Information Standards

This paper is not the appropriate place to present a comprehensive list of existing STDI standards, but some examples are given in Table 2 to demonstrate their variety and how they are formalized (See Section 5.1).

5 The Maturity of Today's Standards for S&T Data and Information

The examples given in Table 2 reflect many aspects of the maturity of today's STDI standards with respect to characteristics such as the formality of the procedure used to

produce the standard, the type of organization in charge, the motivation, the organizations providing the motivation, and the coverage, the robustness and the present-day use. Each is discussed briefly below.

5.1 Formality Used to Develop the Standards

Standards can be developed by a variety of process reflecting different levels of formality.

- *Formal* standards from bodies with processes involving a controlled process emphasizing consensus, recording votes, handling of disagreements, resolving all issues before approval, and balancing among participating organizations
- *Informal* standards from bodies with processes featuring the coalescence of common interest without strict guidelines for approval, participation or issue resolution
- *Implicit* standards from groups with processes as controlled by a small group of people or an individual organization motivated to advance a standard
- *Proprietary* standards from an individual organization that owns and closely controls the standard

Any of these processes may be international, national, discipline-oriented or organization-oriented. Many industrial standards concern the production, characteristics and use of products and services provided by one group for the benefit of another. Because STDI standards are mostly used by the community that developed them, the occurrence of informal or implicit standards is quite common.

5.2 Types of Organizations Developing Standards

As shown in Table 2, the types of organizations developing STDI standards range from formal international standards developing organization such as the International Standards Organization (ISO) [24], less formal international bodies such as W3C [25], national standards organizations, international scientific unions, discipline-oriented bodies, small groups of experts, government agencies, companies and interested individuals. The type of body does play a significant role in the robustness and adoption of the standard. Standards that have been fully vetted by a wide range of interested parties almost always find greater acceptance than those done by a small group of interested persons. The robustness of the former approach also is usually evident as a wider set of views leads to consideration of subtleties and complexity often not recognized by persons or groups espousing with one point of view.

Table 2. Examples of S&T Data and Information Standards

Specific Discipline	Brief Description	Type of Standard (See Text)
Crystallography	Formats [4] developed by an International Scientific Union; covers many areas of crystallography; patented	Informal and Proprietary
Chemical Nomenclature	Four entirely different approaches to chemical nomenclature.	
	CAS Registry Numbers [5] are proprietary, but widely used by others.	Proprietary
	International Chemical Identifier (InChI) [6]; from a small group of interested people, more recently taken over by a Scientific Union	Implicit
	Chemical Mark-up Language (CML) [7]; the first scientific mark-up language; developed informally by a very small group of people; registered with W3C, the standards body of the World Wide Web	Implicit
	IUPAC chemical nomenclature [5] developed under a Union auspice.	Informal
Materials and their Properties	Materials Mark-up Language (MatML) [9]; registered with W3C	Informal
Surface Analysis	ISO standards [10] for collecting surface analysis data	Formal
Digital Imaging in Medical	A variety of standards for various medical imaging techniques developed by an informal group [11]	Informal
Ecology	Ecology Metadata Language (EML) [12]; concepts essential for describing ecological data	Informal
Biodiversity	Many standards for biodiversity collection data [13]	Informal and Implicit
Social and Behavior Sciences	Data Documentation Initiative (DDI) [14]; international mark-up language standard for datasets in the social and behavioral sciences	Informal
Gridded Population of the World	Human population data in a common geo-referenced framework [15]	Informal

Specific Discipline	Brief Description	Type of Standard (See Text)
Earth and Solar Sciences	Standards of the World Data Centers for solar, geophysical, environmental and human dimensions data [17]	Informal and Implicit
Geospatial Information	Geospatial Information Standards (GIS) [18] developed by government, national and international committees	Informal
Seismology	Standard for the Exchange of Earthquake Data [19]	Informal
NASA Satellite-collected Data	NASA Common Data Format [20] is used for many different types of observations	Informal
Engineering Data	ISO Standard for the Exchange of Product Data [21] for sharing data among engineering software used in all stages of a product, from initial design, through manufacturing to use to final disposal	Formal
Information Resources	Dublin Core [22] addresses the metadata about informational resources	Formal
Information Retrieval Services	ANSI Z 39.50 [23] protocols for information retrieval application services (automated catalogs)	Formal

5.3 Motivating Organizations for Standards Development

The motivation for STDI standards still has a firm economic foundation. These standards can save time, money and resources if adopted and used. Yet scientists, especially in academia, government agencies and non-profit organizations, often are not motivated for economic reasons. Instead, one finds other motivations including the desire for intellectual control, mandates to operate data and information collections, or the shear joy of intellectual challenge. The lack of an economic metric – actual savings desired by an organization – often leads to slow progress, especially if competing viewpoints are difficult to resolve. In the case of genomics research, in contrast, the economic incentive of significant research funding for sequencing has provided real motivation for using fairly arbitrary STDI standards in this area.

The initial motivation for STDI standards can come from any number of organizations: private companies, research groups, government agencies, scientific groups and unions, and even standards developing organizations themselves. When

one group dominates the process, the resulting standard can receive a less than enthusiastic reception. Even more challenging is when a standard developed under one auspice is moved to a broader context, such as what has happened with the GIS standards developed by the Federal Geographic Data Committee. Here a government-led group developed a standard that appeared to meet their needs. Then as the standard progressed to broader use, first a national committee and then internationally, ISO became involved. Many additional changes were requested and needed, contrary to the expectations of the initial committee. [25] Indeed once an ISO committee was established, a number of specialized communities added specific extensions, such as for geographically-located biological information.

5.4 Coverage, Robustness and Use of STDI Standards

Most STDI standards have a fairly limited coverage, usually confined to a single sub-discipline. Yet as science deals with more complex aspects of nature, the need for encompassing and overarching standards grows, with specializations or extensions for individual sub-disciplines. Today some attempts are underway – GIS, chemical and materials descriptions, and earth observations. The coming years will require much more effort along these lines.

The robustness of STDI standards, namely the metadata range, flexibility, handling of multiple nomenclatures, and multiplicity of test methods and observation techniques, has been limited to a great extent. As an example of the detail often required for just one set of data, certain methods for testing the properties of composite materials can require several hundred pieces of metadata for full description. Clearly that level of detail requires major effort on the part of the STDI standards developers and present true challenges in making these standards have the necessary robustness. One result is that some standards have three components: a base core of data and information that must be recorded or reported; coverage of data and information that is optionally recorded or reported; and the capability for extension for data and information that in the future could be recorded or reported.

Use of STDI standards today ranges from very limited to quite widespread. For example, CML [7], though intellectually quite advanced, has not received broad acceptance within the chemical data and information community. In contrast, major efforts to create mega-data archives – the International Virtual Observatory Alliance [26] and the Human Genome project [27] have created an environment for widespread adoption and use of STDI standards. The purpose of these efforts is not to collect data but to facilitate research and development and scientific discovery. That goal seems to provide motivation to the scientists for accepting the standards.

6 Barriers to Greater Progress

The advent of mark-up language technology, the emergence of real resources for standards development, driven in part by the push for major data archives, and the greater appreciation of the problems for dealing with the semantics of STDI will bring about new advances in STDI standards. Yet some real and potentially significant barriers remain, including nomenclature, linguistic, sociological and economic, and technical, as discussed in the following sections.

6.1 Nomenclature Barriers

Every area of science and technology has multiple nomenclature systems that arise from historical circumstances based on geography, education, scientific polarization, different languages, conceptual differences and scientific rivalry. In many situations, the different systems are deeply entrenched and not easily reconciled. STDI standards require a common nomenclature, or at least a unique mapping among nomenclatures. Every scientist has her or his story of difficulties in communicating with colleagues using a different nomenclature system, and these stories illustrate the difficulties in resolving the differences. The ideal situation is to invent a new, unambiguous nomenclature, based on a detailed data model that acts as a neutral nomenclature.

The difficulties in integrating historical approaches to the description of objects, the measurement of properties and the independent variables and context for property measurements is greatly exacerbated by the fact that our knowledge evolves over time. What is appropriate to describe a system – chemical, ecological, cellular, whatever – at one time becomes obsolete as new understanding is developed about the true complexity of these systems. A similar evolution occurs in the knowledge about the independent variables needed to describe the properties of a system. For example, experiments done in the early days of high temperature superconductors concentrated on and reported only a few obvious variables. Today the manufacture and preparation of such materials is reported in minute detail with a corresponding explosion of variables reported.

Nomenclature issues get magnified when data are used outside the field of generation, wherein users lack the detailed understanding of all relevant metadata commonly used by experts in the field. In addition, the collision of nomenclature between that of the data generator/reporter and data user often leads to the evolution of new or “Creole” nomenclature, similar to the Creole languages that arise in society. [28] These mixed nomenclatures come from mixing nomenclature in two related areas into something distinct from either of them. For example, the nomenclature of quantum chemistry mixes chemical bonding nomenclature with that of atomic and molecular physics. Individual words may be used in all three areas, but the meaning of the term in the “mixed” field (here quantum chemistry) may be quite different from the two core disciplines.

Finally, the most powerful tools in developing clean nomenclatures – ontologies and data modeling – are often ignored or unknown to STDI standards developers. Ontologies and data modeling, using proven technology and tools, allow for identification of ambiguity, fosters clarity of definition and resolution of the lack of specificity. For example in biology, taxonomies and systematics flow naturally from ontologies and data models, and existing taxonomies and systematics can be refined and improved. Even when ontology and data modeling approaches are known, the time and resource requirements make them easy to ignore. As the stakes for STDI standards grow with the size of time and monetary investments in major data collections, the need for data models to resolve nomenclature problems will become more evident.

6.2 Linguistic Barriers

While we are all intuitively aware that our everyday languages change over time (see the discussion in [28]), the evolution of scientific and technical languages seems to make little impression on scientists and technologists. Yet these evolutionary trends are clearly present in every area of science. The changes are based on exactly the same factors as are involved in everyday languages. A catchy phrase becomes an everyday expression. A concept is developed at one institution and moves slowly or quickly to other institutions, changing slightly with each move. [29] Our understanding of S&T concepts grows, and we need more detailed words and phrases to describe our understanding. People move; words and writing move; and our S&T language evolves as a result.

S&T language evolution dramatically impacts STDI standards: Standards summarize a consensus approach to knowledge, and if our knowledge changes, the standards must change. Languages are dynamic, however, and standards are static. STDI standards developers must recognize that S&T language changes, and standards must contain mechanisms to change with those changes.

6.3 Sociological and Economic Barriers

Typical scientific practice can hinder standards development. Competitiveness, the striving for uniqueness, searching for the unknown, a reluctance to repeat past experiments, and the desire to use new techniques are all manifestations of this situation. Each of these aspects of science works against standards. It is not uncommon for a scientist to say the she or he will use a standard as long as it is based on her or his sense of the state-of-the-art. *My way should be the standard way!*

The lack of economic motivation magnifies this problem. There are very few monetary or prestige awards for using standards in science. Even the possibility of saving time often is unappealing to S&T data generators and collectors, who are

frequently graduate students and postdoctoral researchers trying to do world-class research rather than save time or money.

6.4 Technical Barriers

Today's science and technology are becoming more complex. Science has moved from reductionism to constructivism; the goal of science is moving from discovering the fundamental laws of nature to using those laws to explain real systems that contain enormous numbers of objects; properties of the 6 billion individual people; the virtually countless objects in the universe, biological and terrestrial systems containing 10^{23}-10^{28} molecules; our species of flora and fauna that numbers into the millions, if not tens of millions and more. The complexity of STDI follows the complexity of nature. Real systems are complex; they contain a large number of components, and their properties are subject to a large number of variables. The resulting data and information is complex; and many details must be specified. STDI standards then must address the complexity and reflect the details needed.

A major problem facing STDI standards developers is that these problems face uneven scientific and technological advances. Different groups are interested in different ranges of variables. As a result, they explore variable space unevenly; deeply in some parts, poorly in others. Consequently as STDI standards are developed, they are concentrated in areas of R&D activity, which may not cover all aspects of the systems. Then as new phenomena are found, the standards are inadequate to cover the new areas of interest and require extension and modification. The usual lack of data models means that significant changes are required that often dramatically affect existing standards.

7 The Internationalization of S&T Data and Information Standards

Another significant barrier to progress is the need to accommodate the international nature of today's science and technology. As a result, all the previously identified barriers are additionally exaggerated by the need to account for international activity. Even something as simple as getting international consensus among colleagues with respect to standards is greatly impacted by air travel, language barriers, especially related to detailed nuances, and motivation. This is especially true for bodies such as ISO where standards can take two to five years to get accepted, even with motivated developers.

Fortunately two major forces to overcome the international barriers have emerged. The first is the emergence of the Internet/world wide web/and e-mail. This

connectivity makes international collaboration on STDI standards not only possible, but also much easier. The second is the realization that observational data can easily be lost forever without the necessary precautions. These observational data are *absolutely* non-reproducible. Today's technology seems to present the possibility that these data can be preserved forever, and standards are a key. Activities such as the International Virtual Observatory and biodiversity collaborations are clear indicators of the power of international science.

In June 2005, The National Science Foundation/National Science, Technology, Engineering and Mathematics Digital Library & International Council for Science Unions Committee on Data for Science and Technology Workshop on International Scientific Data, Standards, and Digital Libraries was held at the 5th ACM/IEEE Joint Conference on Digital Libraries. The workshop [30] examined successful models in the development of international standards for languages and tools in use with scientific and technical information. A common theme across the talks reiterated the position that an enormously powerful opportunity now exists to advance scientific endeavor more rapidly through shared access to scientific data both within and across scientific domains.

8 Improving Progress on STDI Standards

None of the barriers identified above are fundamentally fatal to the STDI standardization effort. The barriers are real, and they require considerable effort to overcome. Real progress, however, is possible, especially with knowledge of what the barriers are and why they arise.

Probably the most important factor to improving progress on STDI standard today lies with the need to realize that these standards will constantly evolve as a natural course of events. STDI standards must be designed to allow for growth, both in syntax and semantics. When growth occurs, the standards can change in an orderly manner. Most everyone recognizes that many aspects of computerized data and information change dramatically over time, including hardware, storage media, media formats, database technology, archiving technology, and much more. The semantics of STDI changes equally rapidly. The sooner we build our standards to accommodate those changes, the better the standards will be, and greater adoption and use will follow.

References

1. See the specification for reporting x-ray determination of protein crystal structure determinations: http://deposit.rcsb.org/depoinfo/instruct_xray1.html and Berman, H., The Protein Data Bank: Nucleic Acids Research, 28 (2000) 235-242
2. For example, in recent years protein structures are also being determined using NMR techniques, and new data reporting requirements are being developed; see http://deposit.rcsb.org/depoinfo/instruct_nmr1.html
3. Shoshani, A., Olken, F., Wong, H.: Data Management Perspective of Scientific Data. In Glaeser, P.: The Role of Data in Scientific Progress. Proceedings of the 9th CODATA International Conference. North-Holland, Amsterdam (1985).
4. See the International Union of Crystallography web site:www.iucr.org
5. See the Chemical Abstracts web site: http://www.cas.org/EO/regsys.html
6. See the IUPAC International Chemical Identifier (InChI) web site: http://www.iupac.org/projects/2004/2004-039-1-800.html
7. See the Chemical Markup Language web site: http://wwmm.ch.cam.ac.uk/moin/ChemicalMarkupLanguage
8. See the IUPAC web site: http://www.iupac.org/divisions/VIII/index.html
9. See the MatML web site: http://www.matml.org/
10. See the web site of ISO Technical Committee 201: www.iso.com
11. See the Digital Imaging and Communication in Medicine web site http://medical.nema.org
12. See http:/knb.ecoinformatics.org/software/eml/
13. See links at the GBIF web site: http://www.gbif.org/links/standards
14. See the Data Documentation Initiative web site: http://www.icpsr.umich.edu/DDI/users/index.html
15. See links on the web site: http://beta.sedac.ciesin.columbia.edu/gpw/#
16. See http://www.icpsr.umich.edu/access/dataprep.pdf
17. See links contained on the World Data Center System web site: http://www.ngdc.noaa.gov/wdc/
18. See http://www.opengeospatial.org/specs/?page=specs
19. See http://www.iris.edu/manuals/SEED_chpt1.htm
20. See the NASA Common Data Format web site: http://cdf.gsfc.nasa.gov/
21. See the ISO STEP web site: http://www.tc184-sc4.org/
22. See the Dublin Core Metadata Initiative web site http://dublincore.org/ ANSI Z39.50 (2003)
23. See the International Standards Organization web site www.iso.org
24. See the web site on FGDC/ISO Metadata Standard Harmonization: http://www.fgdc.gov/metadata/whatsnew/fgdciso.html
25. See the International Virtual Observatory web site: http://www.ivoa.net/Documents/latest/
26. See the instructions for submission of gene sequence data: http://www.ncbi.nlm.nih.gov/projects/geo/info/soft2.html
27. McWhorter, J.: The Power of Babel: Henry Holt and Company, New York (2001)
28. Kaiser, D.: Physics and Feynman's Diagrams, American Scientist 93 (2005) 156-165
29. See http://scimarkuplang.comm.nsdl.org/cgi-bin/wiki.pl?SeeAgenda