

# Analysis and Synthesis of Metadata Goals for Scientific Data

**Craig Willis**

*Metadata Research Center, School of Library and Information Science, University of North Carolina, Chapel Hill, Chapel Hill, NC. E-mail: craig.willis@unc.edu*

**Jane Greenberg**

*Metadata Research Center, School of Library and Information Science, University of North Carolina, Chapel Hill, Chapel Hill, NC. E-mail: janeg@email.unc.edu*

**Hollie White**

*J. Michael Goodson Law Library, Duke University, Durham, NC. E-mail: hollie.white@law.duke.edu*

The proliferation of discipline-specific metadata schemes contributes to artificial barriers that can impede interdisciplinary and transdisciplinary research. The authors considered this problem by examining the *domains*, *objectives*, and *architectures* of nine metadata schemes used to document scientific data in the physical, life, and social sciences. They used a mixed-methods content analysis and Greenberg's (2005) metadata objectives, principles, domains, and architectural layout (MODAL) framework, and derived 22 metadata-related goals from textual content describing each metadata scheme. Relationships are identified between the domains (e.g., scientific discipline and type of data) and the categories of scheme objectives. For each strong correlation ( $>0.6$ ), a Fisher's exact test for nonparametric data was used to determine significance ( $p < .05$ ).

Significant relationships were found between the domains and objectives of the schemes. Schemes describing observational data are more likely to have "scheme harmonization" (compatibility and interoperability with related schemes) as an objective; schemes with the objective "abstraction" (a conceptual model exists separate from the technical implementation) also have the objective "sufficiency" (the scheme defines a minimal amount of information to meet the needs of the community); and schemes with the objective "data publication" do not have the objective "element refinement." The analysis indicates that many metadata-driven goals expressed by communities are independent of scientific discipline or the type of data, although they are constrained by historical community practices and workflows as well as the technological environment at the

time of scheme creation. The analysis reveals 11 fundamental metadata goals for metadata documenting scientific data in support of sharing research data across disciplines and domains. The authors report these results and highlight the need for more metadata-related research, particularly in the context of recent funding agency policy changes.

## Introduction

Metadata for the representation and description of scientific data are an essential component of contemporary scientific communication. Over the past several decades, communities from within the physical, life, and social sciences have developed metadata schemes to facilitate the documentation, exchange, archiving, and reuse of research data. Many of these developments are associated with discipline-specific data repositories. Obvious positive outcomes stemming from community-driven metadata practices include discipline support for science beyond a single lab and data sharing across an entire community. Despite noted benefits, the proliferation of discipline-specific metadata schemes has also contributed to establishing artificial barriers to data discovery and reuse across disciplines. These barriers, frequently associated with metadata semantics and data structures, interfere with scientific progress along multidisciplinary, interdisciplinary, and transdisciplinary lines. Together the barriers can interfere with progress supporting our contemporary understanding of science.

Contemporary science has been characterized as a combination of specialization and multidisciplinary or interdisciplinary research, with growing numbers of specialties developing out of the boundaries between disciplines (Garvey, 1979). The growth of cross-boundary research in

---

Received November 30, 2011; revised February 1, 2012; accepted February 19, 2012

© 2012 ASIS&T • Published online 26 June 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.22683

the sciences is traced more thoroughly by interdisciplinary scholars such as Klein (1999) and Hubenthal (1998). Klein (1999) provides insight by highlighting problem-solving approaches pursued by researchers “crossing boundaries” and exchanging information, techniques, and tools with scientists in other domains. Scientific interdisciplinarity moves beyond tools and techniques to form domains, such as ethnobiology and biochemistry, specifically created to represent multiple perspectives (Hubenthal, 1998).

As new specialties emerge, artificial barriers impeding scientific progress need to be eliminated. This goal is becoming more prevalent with the accelerated growth of digital data, and new opportunities presented by networked technology. Recent attention to preserving and providing access to digital data can be seen as a call for research directed at eliminating data silos. Examples include the following:

- An adamant call for greater attention to the diversity of data following the long tail of science (Heidorn, 2008)
- An acute need to globally manage the deluge of digital data (e.g., Hey & Trefethen, 2003; National Science Board, 2005)

National and international policies and calls for an infrastructure supporting data sharing (e.g., NSF Data Sharing Policy, <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>; NIH Data Sharing Policy, [http://grants.nih.gov/grants/policy/data\\_sharing/](http://grants.nih.gov/grants/policy/data_sharing/); Committee on Science, Engineering, and Public Policy (US), and Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, 2009; NSF Task Force on Cyberlearning, 2008; DCC/JISC, 2008).

These calls, the grand challenges they highlight, and the need for metadata informed the research presented in this article. Open, shared, and interoperable metadata systems can help effectively preserve and provide access to scientific data across disciplines. Historical data documentation practices are manifest in the wide array of discipline-focused metadata standards. Understanding the scope and expanse of metadata systems that already exist is a necessary first step in this area. Research needs to identify metadata practices that cut across domains and that should be applied universally. An inquiry into metadata practices and goals can contribute to the pool of knowledge for supporting a more interoperable environment—one that is hospitable to interdisciplinary and transdisciplinary science.

This article presents results from a systematic analysis of metadata schemes for scientific data, with the goal of understanding the scope and extent of existing metadata practices. The research was conducted to identify (a) existing metadata practices that are common across disciplines and (b) unique metadata practices within disciplines that may be more widely applicable. A mixed-methods content analysis (Spurgin & Wildemuth, 2009; Zhang & Wildemuth, 2009) was conducted to examine the principles, objectives, domains, and complexity of nine metadata schemes currently used for representing scientific data. The metadata objectives, principles, domains, and architectural layout

(MODAL) framework (Greenberg, 2005) was instrumental in guiding this study; MODAL is a model designed for studying the overall goals, scope, and effectiveness of metadata schemes.

This article is organized as follows: The next section reviews different types of scientific data and metadata-driven goals. Research approaches and frameworks for studying metadata are then presented, with focused attention given to Greenberg’s MODAL framework. The study’s research questions, methods, sample, and procedures are then detailed. The results and context discussion follow. The final section includes a conclusion and considers next steps toward a better understanding of metadata needs for scientific data and the elimination of data silos.

## Scientific Data and Metadata Goals

Understanding the range and varied types of scientific research data is important for exploring the population of metadata schemes developed for scientific data. Scientific data vary greatly, as do their intended and potential use over time. This variance is reflected in the metadata schemes supporting data use and manipulation. In exploring this topic, we first provide an overview of the range and types of scientific data. The overview is followed by an examination of metadata-driven goals underlying specific types of scientific data.

### *Types of Scientific Data*

Scientific data can be classified in a variety of ways. Data can be grouped by the discipline that creates and plans to use the data (e.g., chemistry data or social sciences data) or the data collection method (e.g., survey data or data from computer models). Classifying types of scientific data can be helpful for understanding the similarities and differences as well as the intended and potential use of data over time. The US National Science Board (NSB; <http://www.nsf.gov/nsb/>), the United Nations Educational, Scientific, and Cultural Organization (UNESCO; <http://www.unesco.org/>), and the International Council for Science Committee on Data for Science and Technology (CODATA; <http://www.codata.org/>) are substantial organizations that have presented well-understood classifications for scientific data. Their work informs the research presented in this article.

In a review of digital data collections, the NSB (2005) classifies digital scientific data based on origin: whether observational, computational, or experimental. According to the NSB, observational data cannot be recollected and are archived indefinitely. Data that is the result of computer models or simulations can be reproduced if adequate information is provided about the computer hardware, software, and inputs. Experimental data can often be reproduced, although there are cases where experimental conditions or variables are unknown. Based on these categories, it is apparent that the requirements for managing and describing research data differ depending on how the data was collected or generated.

TABLE 1. Lide's (1981) three classes of scientific data.

Class A	Repeatable measurements on well-defined systems	In principle, data are subject to verification by repeating the measurements in different laboratories at different times.
Class B	Observational data	Time- or space-dependent measurements that cannot, in general, be checked by remeasurement
Class C	Statistical data	Including nonscientific or nontechnical data

In an extensive study of the problems of accessibility and dissemination of data conducted over a quarter of a century ago by CODATA on behalf of UNESCO, Kotani (1975) developed a comprehensive classification of scientific data. The classification consists of 15 distinct categories or facets, such as “data which can be measured repeatedly” or “data which can be measured only once,” “location independent” or “location dependent,” “primary” or “derived,” “determinable” or “stochastic,” etc. The complete set of categories with examples is included in the Appendix. Lide (1981) collapses the UNESCO/CODATA categories into the three broad classes—repeatable measurements on well-defined systems, observational data, and statistical data—presented in Table 1. These classes parallel those defined by the NSB.

Repeatable measurements on well-defined systems result in data that can be verified independent of time or location, as long as the procedures and relevant variables are sufficiently documented. The NSB (2005) refers to this type of data as experimental research data, distinguishing it from observational data or data produced by computer models. Experimental data may be associated with a particular methodology or instrument (e.g., x-ray crystallography). Disciplines such as physics, chemistry, or thermodynamics are generally associated with this type of data, although disciplinary boundaries are not distinct.

Observational data, compared to most experimental data, cannot be recollected, remeasured, or verified. Data are typically time and/or location dependent. This context is set by the fact that much of the value of observational data is in its secondary analysis. Examples are offered by Kelling's (2008) and Michener's (2006) descriptions of observational data collection and reuse in the biodiversity and ecological research communities. In both cases, the observational context, including time, location, and method of collection, are essential to facilitating secondary reuse and analysis, which may occur long after the original study. Disciplines frequently associated with observational data include ecology, biology, and the social sciences. Subfields of biology, such as molecular or structural biology, are more closely associated with experimental research data.

Statistical data, computational models, and simulations can also be recreated and verified, as long as sufficient

information about the original process is captured. This includes details about the inputs, software, or instruments used. Like observational data, statistical data is also subject to reuse and possible transformation.

The CODATA glossary (Westbrook & Grattidge, 1991) includes definitions for these three broad classes of data with additional qualifications. Experimental data is further defined as data “gathered from an experiment and before evaluation or other data validation techniques have been applied.” Through this definition, experimental data is differentiated from evaluated and validated data, which have been subject to validation processes or shown to be generated according to standard methods. The CODATA glossary further distinguishes raw data, or data that has not been processed from its original form. In fields such as thermodynamics or crystallography, evaluated or validated experimental data are generally of more interest than the raw experimental data.

The broad classes of scientific data and their varied contextual origins suggest that metadata will have different features related to each class. This expectation motivates the current study. Because the classes of scientific data span across disciplines depending on research methodologies, the features of associated metadata schemes should as well. We further explore this expectation in the next section by looking at goals underlying metadata schemes targeting specific types (classes) of scientific data. A chief motivator is to capture a picture of the metadata environment supporting scientific data. A better understanding of this environment is necessary for further developing a robust data documentation infrastructure, particularly for an environment that is more hospitable to interdisciplinary and transdisciplinary research.

#### *Metadata Goals for Scientific Data*

Metadata, a fundamental component of any information system, is goal driven. Metadata-related goals are shaped by both (a) the type of resource being represented and frequently stored in an information system and (b) the desired uses of the represented resource (Day, 1999; Greenberg, 2003, 2009). It follows that a thorough study of metadata for scientific data would involve examining goals related to specific types of scientific data. For instance, one might take Lide's (1981) definitions, given in the above section Types of Scientific Data, as a starting point for determining the goals or objectives of metadata for each class of scientific data. It is reasonable to propose that communities engaged in experimentation require metadata documenting context and procedures. This information is crucial for independently verifying the experiment, and likely of greater immediate importance than data preservation and archiving. Communities responsible for observational data require documentation of the observational context as well as preservation of the resulting data for long-term reuse. These observations make sense, although recording goals specific to “type of data” fails to capture goals that differ across

disciplines. Moreover, there seems to be little analysis of metadata goals relating to specific types or classes of data.

Despite the above noted limitations, there is a rich body of literature supporting metadata work in various scientific domains, and this literature gives insight into metadata related goals that are in fact applicable across disciplines. This approach displays aspects of what Hjørland and Albrechtson (1995) call “domain analysis,” getting a sense of community-specific needs.

Researchers from various scientific fields have published exemplary works articulating community needs that have informed discipline-oriented metadata standards. A few examples follow here:

- Hall, Allen, and Brown (1991) are known for articulating data description needs in the crystallography community in their call for a “general, flexible, rapidly extensible, and universal file format.” Viewed within a more general framework, their work articulates needs applicable to all sciences. They underscore the central goals of crystallographic data description, noting archiving data and support for the exchange of data between different software packages, laboratories, and authors and journal publishers. They present the Crystallographic Information File (CIF), a format that is central to the Cambridge Structural Database (CSD), a highly successful repository of scientific data today.
- Frenkel et al. (2006) present goals for the thermodynamics research community that are similar to those articulated for the crystallography community. They describe a global process for exchanging thermodynamics data, facilitated by a standard metadata format. Goals described by the authors include interoperability between software packages and organizations, archiving data, and exchanging data with full support for data provenance including details about the conditions under which the data were generated.
- Michener (2006) and Jones, Berkley, Bojilova, and Schildhauer (2001) are well known for articulating the goals of the ecological research community, also similar to those described by Hall et al. (1991) and Frenkel et al. (2006). According to Michener, there are three central goals supported by metadata for scientific data in ecology: (a) increasing the longevity of data, (b) increasing the reuse of data, and (c) facilitating sharing of data. These goals are achieved, in part, by capturing sufficient details about the research context and structure of the data to support long-term reuse.

The work outlined above and other community-specific work (e.g., Brazma et al., 2001; Ryssevik & Musgrave, 2001; Spellman et al., 2002; Westbrook & Bourne, 2000) form a collective body of literature on the metadata-driven goals for scientific data. Despite this observation, there is little research identifying and assembling these goals in a uniform way. Even so, it is clear that this type of analysis has the potential to reveal a set of universal goals applicable across disciplines, such as archiving and data exchange. The need for a more universal metadata framework stems from national and international attention to address the growing digital data deluge (Hey & Trefethen, 2003; NSB, 2005). At this time, however, most reports and recommendations are

not backed with empirical evidence documenting common or universal metadata-driven goals. More metadata-focused research is needed to move toward an increasingly interoperable environment where scientific data is to be shared across domains and communities, and existing data silos are eliminated. Specifically, research is needed to examine the relationship among metadata goals articulated by communities, which is the central purpose of the study presented in this article. By understanding the similarities and differences within and among schemes, the intent is to identify areas where metadata practices contribute to the creation of artificial barriers to data sharing. This step is necessary to eliminate these barriers—an essential step for improving interoperability and to creating an environment that is more hospitable to interdisciplinary and transdisciplinary research. Study in this area also requires a framework to aid the analysis. The next section considers metadata models and introduces the MODAL framework, which guided the analysis reported on herein.

### Studying Metadata: Approaches and Frameworks

The growth of digital resources has resulted in a proliferation of metadata schemes, prompting analysis to understand the similarities and differences between schemes. The most popular method for comparison is crosswalk analysis, when various schemes are mapped out, property-by-property (Chan & Zeng, 2006). This approach, likely the most economical in terms of time demand, allows for one to see which metadata schemes are useful for an initiative, and where potentially new properties are needed or schemes might be combined.

Existing comparisons of metadata schemes on functional or contextual levels seem quite limited. Scheme creators are more prone to adopt or modify an existing standard or possibly create a new one. One explanation for this may simply be the nature of metadata. There is generally a pressing need to get a system up and running or supporting a specific task. It may be more efficient to create a new scheme or modify an existing one than to first map out metadata goals and functions. As a result, metadata work is frequently more ad hoc, and not necessarily supported by comparative analysis.

Another reason for limited comparisons may be the absence of frameworks to guide analysis. Several frameworks have been developed to guide metadata work, and assist with the packing, use, and reuse of metadata. Examples include the Functional Requirement for Bibliographic Records (FRBR; <http://www.ifla.org/publications/functional-requirements-for-bibliographic-records>) for bibliographic data; Metadata Encoding & Transmission Standard (METS; <http://www.loc.gov/standards/mets/>) for partitioning and packaging different types of metadata in the digital library/repository environment; and the Singapore Framework (<http://dublincore.org/documents/singapore-framework/>) for developing application profiles, integrating, and using metadata properties across domains.

These frameworks provide a useful means for comparing metadata semantics at different levels, but do not address metadata goals. The Singapore Framework identifies functional requirements as a mandatory component of an application profile, but does not provide a means to compare requirements and goals among schemes.

Valuable and influential frameworks have been developed to support the assessment of the quality of metadata and information in general (Bruce & Hillman, 2004; Margaritopoulos, Margaritopoulos, Mavridis, & Manitsaris, 2008; Stvilia, Gasser, Twidale, & Smith, 2007). The goals and objectives of metadata schemes might serve as the subject of quality analysis (e.g., do the goals meet the needs of the community) or as a factor in the evaluation of the scheme implementation (e.g., does the implementation meet the goals stated by the community). In this current study, we focus on metadata scheme goals and objectives. This work may provide a foundation for later integrating frameworks oriented toward metadata quality.

To move forward with our analysis, we elected to work with Greenberg's (2005) MODAL framework. MODAL provides a framework for the analysis and comparison of the fundamental aspects of metadata schemes that are important to understanding the current environment. Metadata scheme objectives are commonly articulated in natural language in the introductions of scheme specifications or published articles describing the scheme. Following MODAL, objectives "identify the overall aims and goals of the scheme" and help us to understand factors motivating a community's scheme creation efforts. The MODAL framework defines three domains for the analysis of metadata schemes. The *environmental domain* identifies the discipline or community responsible for the scheme. The *object class domain* identifies the type of thing described by the scheme. Examples include events, persons, places, or information resources. The *object format domain* identifies the format of the object being described. Examples include geospatial materials, multimedia, digital-like objects, or books.

The MODAL framework also captures structural information about the design of a scheme, or its architectural layout. Examples of structural types include the extent (or number of elements), granularity of elements, simple/compound elements, and cardinality.

MODAL, when first published, was said to be an initial rendering, and Greenberg noted it was "unfinished." We (including the author of MODAL) recognize that there are limitations with the first rendering of MODAL. Despite this predicament, MODAL provides a very useful place for initiating research specifically on metadata goals and functions. Research has not revealed another model that is robust enough for facilitating needed research in area. Moreover, the MODAL framework has gained a fair amount of attention in recent research literature (Bountouri & Gergatsoulis, 2009; Lim & Chiew, 2011), confirming its usefulness as an evaluative tool. In sum, compared to the models referenced above, the MODAL framework offers the most explicit

means for substantial study of metadata goals, which is the purpose of this research article.

### Research Questions

The proliferation of metadata schemes for documenting and representing scientific data contributes to artificial barriers in discovery and reuse disciplines and domains. A better understanding of the metadata goals articulated by individual communities will help to define a more universal approach to describing scientific data. This is our chief objective in this article. The questions guiding this study are:

1. What is the scope of scientific metadata schemes?
2. What are the similarities and differences between scientific metadata schemes?
3. What may be the fundamental requirements of metadata schemes for scientific data?

The procedures for data collection and analysis are detailed below.

### Method and Procedures

This study uses a qualitative directed content analysis approach to examine metadata scheme objectives, principles, and domains. In directed content analysis, coding starts with a preexisting theory or prior research findings (Zhang & Wildemuth, 2009). We use a basic codified approach initially outlined by Krippendorff (1980). The MODAL framework provides the preexisting categories and theory. Quantitative content analysis is used for the analysis of scheme architectural layout.

### Data Collection

To investigate the research questions, we selected a sample from a set of over 50 metadata schemes used to describe and document scientific data in the physical, life, and social sciences, as well as the three broad classes of experimental, observational, and statistical data. Several constraints were put on the sample:

- The scheme must describe scientific data sets, not merely represent scientific information.
- The scheme must be used in an active scientific data repository.

For example, the chemical markup language (CML; Murray-Rust & Rzepa, 1999), widely used for the representation of chemical structures in documents, is not used for the documentation of scientific data sets or used in a data repository; therefore, it is not included in this study. The schemes are listed in Table 2. These constraints allow us to focus on mature metadata schemes with similar applications across domains that have been adopted by a community and

TABLE 2. Sample of metadata schemes used in this study.

Scheme name	Repository
Crystallographic Information File (CIF)	Cambridge Structural Database (CSD)
Data Documentation Initiative (DDI)	ICPSR, CESSDA
Darwin Core (DwC)	GBIF Data Portal
Ecological Metadata Language (EML)	ESA Data Registry
Macromolecular Crystallographic Information File (mmCIF)	Protein Data Bank (PDB)
MIAME Notation in Markup Language (MINiML)	Gene Expression Omnibus
Micro-Array Gene Expression Markup Language (MAGE-ML)	ArrayExpress
NEXML	TreeBase
ThermoML	ThermoML Archives

have sufficient documentation for content analysis. After applying these constraints, the result is a sample of nine metadata schemes used to describe scientific data in digital repositories.

For content analysis, the sampling process further included the collection of three to six central texts per metadata scheme along with the scheme implementation files (e.g., XSD or DTD). The central texts include technical specifications, user documentation, and published journal articles used to communicate about the scheme to the user community. The content analysis procedures are outlined below.

This study is intended to contribute to further understanding of metadata-related goals and needs articulated by scientific communities and relies on documentary evidence in the form of publications and scheme implementation files. It is possible that the needs and objectives stated by scheme designers are incomplete or do not fully reflect the needs and goals of the target community. The consistency we did find among documentation for each scheme and the supporting articles, however, led us to conclude that the documentation is a valid source for analysis. In other words, as accepted standards, these documents represent a general consensus among community members.

### Data Analysis

The data analysis process is composed of qualitative and quantitative content analysis with pairwise correlations, Fisher's exact and Wilcoxon signed rank tests. The qualitative content analysis was applied in three phases. During the first phase, a set of categories of scheme objectives and principles was generated based on the interpretation of each of the sample texts in the context of the MODAL framework. Analysis was limited to those sections of each text focused on the general description of the scheme. During the second phase, the initial set of categories was further refined through a process of definition and a secondary

reading of the sampled texts. In the third phase, texts were coded based on the defined categories.

Quantitative content analysis was applied through the direct analysis of the scheme implementation files based on the categories defined in the MODAL framework. The scheme encoding, numbers of elements, numbers, and types of files were identified and counted directly.

The resulting data was prepared for import into the JMP (version 9; <http://www.jmp.com>) statistical analysis package. Objectives, principles, and domain data were coded as binary nominal fields with values of 0 or 1 indicating the absence or presence of each category. Architectural layout data were coded as binary nominal and continuous fields. Pairwise correlation was run for all values. For each strong correlation ( $>0.60$ , probability  $< 0.05$ ), a Fisher's exact (for binary nominal) or Wilcoxon signed rank test (for nominal/ratio data) were run to determine significance. The Fisher's exact and Wilcoxon signed rank tests were selected for analysis of nonparametric data. The results of this analysis are reported in the Results section below.

## Results

In this section, we present the results of the analysis of nine metadata schemes for the description of scientific data, using a combination of qualitative and quantitative content analysis directed by the MODAL framework. The selected schemes are representative of the physical, life, and social sciences, and are used to describe experimental, observational, and statistical data sets, including both digital and physical collections.

Schemes include the Ecological Metadata Language (EML) and Darwin Core (DwC), used to describe biological and ecological studies and specimen collections; the Crystallographic Information File (CIF) and Macromolecular Crystallographic Information File (mmCIF), used to describe physical and biological crystallographic structures; the Data Documentation Initiative (DDI), used to describe social sciences data; the Micro-Array Gene Expression Markup Language (MAGE-ML) and MIAME Notation in Markup Language (MINiML), used to describe molecular abundance data; ThermoML, used to describe evaluated experimental thermophysical and thermochemical property data; and NeXML, used to describe phylogenetic trees.

Using the MODAL framework, the schemes were classified based on the discipline that they serve. The results of this classification are listed in Table 3. Six of the nine schemes (66%) represent data from the life sciences (Darwin Core, EML, MAGE, MINiML, mmCIF, NeXML), two schemes (22%) represent data in the physical sciences (CIF, ThermoML), and one scheme (11%) represents data in the social sciences (DDI).

The schemes were then classified based on Lide's (1981) categories of experimental, observational, and statistical data. Eight of the nine schemes (89%) represent data from experimental studies. Three of the nine schemes (33%) represent data from observational studies (including both

TABLE 3. Classification of schemes by domain.

Scheme	Environmental domain	Object class domain	Object format domain
CIF	Physical sciences—Crystallography	Experimental studies	Digital data—crystallographic structures
Darwin Core	Life sciences—Biology	Observational studies Specimen collections	Digital data—biological collections Physical collections
DDI	Social sciences	Experimental studies, observational studies, and statistical studies	Digital data—social science statistical data
EML	Life sciences—Ecology	Experimental studies Observational studies	Digital data—ecological observation and experimental results
MAGE	Life sciences—Molecular biology	Experimental studies	Digital data—molecular abundance data
MINiML	Life sciences—Molecular biology	Experimental studies	Digital data—molecular abundance data
mmCIF	Life sciences—Structural biology	Experimental studies	Digital data—macromolecular crystallographic structures
NEXML	Life sciences—Phylogeny	Experimental studies	Digital data—phylogenetic trees
ThermoML	Physical sciences—Thermodynamics	Experimental studies	Digital data (Thermodynamic properties)

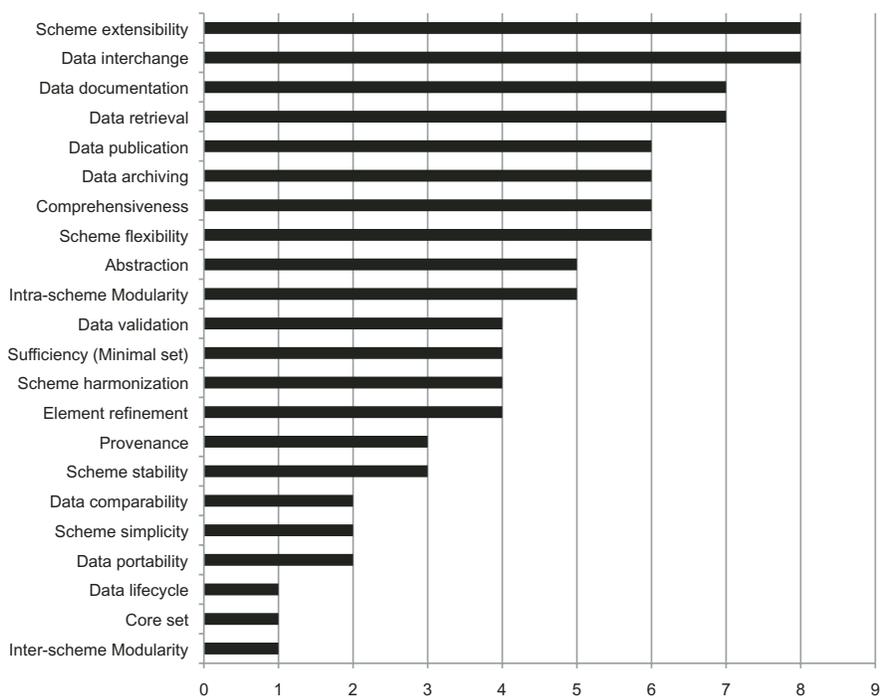


FIG. 1. Frequency of metadata goals/objectives across scheme sample.

digital and physical collections). One of the nine schemes (11%) represents data from statistical studies.

The object format domain for each scheme was identified as either digital data or physical collections. All nine of the schemes are intended to describe digital data sets. Only one scheme, Darwin Core, is intended to also describe physical collections or specimen.

As described in the Data Analysis section, the objectives and principles of each scheme were identified using directed qualitative content analysis on selected texts. During the initial phases of content analysis, a set of 22 conceptual categories was identified based on qualitative readings. A complete list of these categories and their definitions are included in the Appendix. The texts were then coded for the

presence/absence of each conceptual category. The identified categories were totaled across the set of schemes. The results of this analysis are presented in Figure 1. The absence of a category does not mean that the function is absent from the scheme, only that it is not an objective stated by the scheme creators.

No single objective was present in all schemes. Eight of the identified objectives (36%) were present in over two thirds of the schemes. These include data interchange, scheme extensibility, data documentation, data retrieval, data publication, data archiving, comprehensiveness, and scheme flexibility. Examples of these objectives taken from the analyzed texts include phrases such as “interchange/exchange standard,” “extensible format,” “can be modified

TABLE 4. Significant relationships found between scheme objectives and domains.

Objective 1	Objective 2	Correlation	<i>p</i> Value
Scheme harmonization (objective)	Observational data (domain)	0.79	.0476
Abstraction (objective)	Sufficiency (objective)	0.80	.0476
Data publication (objective)	Element refinement (objective)	(0.79)	.0476

to meet domain-specific needs,” “facilitate discovery and retrieval,” “for submitting data to journals and databases,” “archiving standard,” and “a comprehensive dictionary.”

Six objectives (27%) were present in more than one third but fewer than two thirds of the schemes. These include abstraction, intrascheme modularity, data validation, sufficiency, scheme harmonization, and element refinement. Examples of these objectives include “distinction between content model. . . and the syntactic implementation,” “for the validation of data,” “minimal information required,” or “the format encompasses. . . all information a systematist might wish to use.”

The remaining eight objectives (36%) were present in less than one third of the schemes. These include provenance, scheme stability, data comparability, scheme simplicity, and data portability. Three objectives were unique to a single scheme: data lifecycle (DDI), core set (Darwin Core), and interscheme modularity (Darwin Core). Examples of these objectives include “existing items must *never* be changed,” “stable semantic definitions,” “document. . . data across its life course,” and “comparability of studies.”

After coding the domain and objectives data, a pairwise correlation was run for all values. For each strong correlation (>0.6), a Fisher’s exact test for nonparametric data was run to determine significance ( $p < .05$ ).

Although the sample size is small, significant relationships were present between the domains and objectives of the schemes, as listed in Table 4. Schemes describing observational data are more likely to have scheme harmonization (compatibility and interoperability with related schemes) as an objective. Schemes with the objective of abstraction (a conceptual model exists separate from the technical implementation) also have the objective of sufficiency (the scheme defines a minimal amount of information to meet the needs of the community). Schemes with the objective of data publication do not have the objective of element refinement.

During the next stage of analysis, scheme implementation files were inspected to capture architectural layout information including supported encodings, structural types, extent (i.e., the number of elements in the scheme), and levels of hierarchy. The results of this analysis are listed in Table 5.

The scheme encodings were captured in terms of the higher level formats used to define the scheme (e.g., XSD, DTD, or RDF). Six of the nine schemes (66%) supported multiple encodings. Eight of the nine schemes supported XSD encodings (89%). Four of the nine schemes (44%) supported flat-file (CSV or text) encodings. No significant relationships were found between the domain and the encoding.

The structural types were captured by the terms used to describe the structure of the scheme and are taken from the language of the scheme itself (e.g., data blocks, elements, simple types, or complex types). STAR DDL, the higher level encoding used for the CIF and mmCIF, is defined in terms of *data blocks*, *categories*, and *data items*. EML, encoded using XSD, defines its own structural language with terms such as *resource modules* and *supplemental modules*, in addition to the standard XSD terms *element*, *simple type*, or *complex type*.

The extent of the scheme is captured as raw counts of the elements in the scheme based on the scheme-specific structural types. The term *element* is often used to describe the finest-grained components of a scheme (e.g., data items, terms, properties, or XML elements). There are an average of 563 elements per scheme, with a minimum of 142 (MiNIML) and a maximum of 1,802 (mmCIF).

The number of files indicates the number of physical schema files used to specify the scheme. Five of the nine schemes (56%) use more than one file, three of which (33%) use more than 20 files in the scheme definition. There are an average of nine files used per scheme, with a minimum of one and a maximum of 25 (EML).

The number of levels indicates the levels of hierarchy (i.e., depth) supported by the scheme. There are an average of five levels per scheme with a minimum of three (Darwin Core) and a maximum of 10 (MAGE).

Architectural layout information was coded using a combination of binary nominal and ratio data. A pairwise correlation was run for all values in conjunction with the domain and objectives data. Fisher’s exact and Wilcoxon signed rank tests were run to determine significant relationships ( $p < .05$ ). The results are listed in Table 6.

Significant relationships were identified between the encoding and scheme objectives. Schemes that support a flat file encoding do not have comprehensiveness or data archiving as objectives.

## Discussion

The results presented above provide a starting point to help to understand the similarities and differences among existing metadata schemes and provide insight into metadata-driven goals that are applicable across disciplines. These results also present the opportunity to consider the scope of scientific metadata. The following discussion sections cover these aspects, examining similarities and differences, reflecting on apparent universal goals, and considering further the scope of scientific metadata.

TABLE 5. Summary of scheme architectural layout based on MODAL.

Scheme	Encodings	Structural types	Extent	# of files	# of levels
CIF	STAR DDL	Data blocks	18	1 DDL	5
	XSD	Categories	62		
DwC	XSD	Data items	486	10 XSD	3
		Terms/properties	178		
		Classes	9		
DDI	XSD	Elements	797	22 XSD	6+
		Complex types	296		
		Simple types	599		
EML	XSD	Resource modules	4	25 XSD	5+
		Supplemental modules	6		
		Elements	579		
		Complex types	174		
		Simple types	54		
MAGE	DTD	Packages	16	1 DTD	10
	Flat File				
MINiML	XSD	Entities	76	1 XSD	4
		Elements	423		
		Elements	14,224		
mmCIF	STAR DDLXSD	Complex types	19	1 DDL	5
		Data blocks	25		
		Categories	168		
NeXML	XSD	Data items	1,802	21 XSD	5+
		Blocks	8		
		Elements	196		
ThermoML	XSD	Complex types	132	2 XSD	7
		Simple types	71		
		Blocks	4		
		Elements	470		
		Complex types	94		
		Simple types	52		

TABLE 6. Significant relationships between architectural layout features and scheme objectives.

Feature	Objective	Correlation	<i>p</i> Value
ENC – Flat File	Objective—Comprehensiveness	(0.79)	.0476
ENC – Flat File	Objective—Data Archiving	(0.79)	.0476

### *Similarities and Differences Among the Metadata Schemes Examined*

Through the analysis of the documentation describing metadata schemes for scientific data, we are able to construct a set of categories representing goals expressed by scheme creators. The 22 identified categories are discussed in the above Results section and definitions are included in the Appendix. These categories were identified through analysis and refinement based on the textual content describing each scheme. The nine metadata schemes represent a diverse set of disciplines and describe different types of data for communities with distinct approaches to science and communication. Even with these differences, the scheme goals and architectural layouts reflect similarities that are independent of discipline or type of data.

Metadata is part of a larger information ecology that includes systems and software. Metadata is created and consumed by software packages, often requiring the involvement of systems and software specialists at the time of scheme creation. Several of the commonly expressed goals identified in this study are as equally applicable to systems and software as to metadata schemes. Extensibility, flexibility, modularity, and portability are all common concepts found in standard texts on systems and software architecture (Bosch, 2000; Buschmann et al., 1996). These goals are independent of domain.

Other goals common across two thirds of the studied schemes reflect needs specific to repositories, but are still independent of discipline and type of data. These include data documentation, interchange, retrieval, and archiving, which are central functions of any archival information system or repository (Consultative Committee for Space Data Systems [CCSDS], 2009; Higgins, 2008). It is worth looking at cases where a community does not state certain goals.

Darwin Core is unique among the schemes studied. Influenced by the Dublin Core Metadata Initiative (DCMI; <http://www.dublincore.org/>), Darwin Core has been used as a vocabulary to compose other metadata schemes used to describe scientific data, such as the Dryad application profile

(Carrier, Dube, & Greenberg, 2007). This feature of Darwin Core is consistent with the design and application of Dublin Core in digital libraries and closely related to current trends in linked data. Although data exchange, discovery, and retrieval are stated goals, data documentation is not. According to the Darwin Core scheme description, the “occurrence of taxa in nature are documented by observations, specimens, and samples” (Taxonomic Data Working Group, 2009). The purpose of Darwin Core is to represent these observations and specimens and in this sense it is self-documenting. Unlike the other schemes, Darwin Core is not intended to document a set of data, but to represent observations. Darwin Core was only recently adopted as a repository format (Global Biodiversity Information Facility [GBIF], 2010).

Although there are many similarities in scheme goals, there are also notable differences. A common distinction between schemes is evident in the goals of scheme comprehensiveness and scheme sufficiency. Several schemes (CIF, mmCIF, NEXUS, and ThermoML) cite scheme comprehensiveness as a goal, to provide a set of elements that support the thorough description of data in the domain. Others, such as MAGE-ML and MiNiML, cite scheme sufficiency, to support the minimal amount of information needed to document an experiment. Two schemes, DDI and EML, express both goals. These two schemes are intended to be comprehensive, yet support instances of description using a minimal number of required elements.

A related goal is that of scheme simplicity, as expressed in both Darwin Core and MiNiML. Users of metadata schemes often have varying degrees of expertise and access to programmers and IT staff (Rayner et al., 2006). Complex schemes with hundreds of elements and numerous schema files require a level of expertise that is beyond some researchers and small labs. The goal of simplicity is often presented in this light. Some schemes, such as DDI and MAGE-ML, have created simplified versions (i.e., DDI-Lite and MAGE-TAB) to address different needs within their communities. The ability to create both complex and simple versions of a scheme is also related to the goal of abstraction.

Scheme abstraction is the process of creating a separate conceptual model that is independent of a particular scheme rendering. Over half of the schemes in this study include abstract or conceptual models, usually described in prose lists or modeled in diagrams. By defining the scheme in this way, communities are able to create multiple different renderings that meet the same general requirements and support community goals, independent of a particular encoding or format. A good example is MAGE-ML and MAGE-TAB, which both conform to the abstract MIAME goals with very different implementations.

Another notable difference among schemes is in the goal of data publication. Six of the nine schemes explicitly cite supporting publication in journals and data repositories as a goal. The remaining three schemes (Darwin Core, EML, and NEXUS) were designed with purposes other than supporting

publication, but were later adopted as standard formats for data publication.

This situation raises the question of how the origins of a scheme affect our ability to understand and contextualize community goals. For example, CIF emerged from the crystallography community’s practice of auxiliary publication—the storage of supplemental material associated with research articles—and from its inception was part of a process of publishing data associated with research (Hall & McMahon, 2006). Similarly, ThermoML was informed by the historic practice of data compilation in the thermodynamics community and was also intended from inception to support publication of data associated with research (Frenkel et al., 2006). Conversely, Darwin Core began as a Z39.50 profile intended to support search and retrieval in natural history collections and observation databases, similar to Z39.50 for library catalogs (University of Kansas, 2004). Later versions evolved into general-purpose vocabularies for the representation of observations. Only recently was Darwin Core adopted as a standard for the description of occurrence data associated with research publications (GBIF, 2010). EML followed a similar path, beginning as a format for the description of data contained in local data catalogs, only to be later adopted as the standard for documenting datasets in the ESA (Ecological Society of America) archives (Bain & Michener, 2002).

The examples discussed above also raise the question of the distinction between databases, data banks, catalogs, archives, and repositories across disciplines. Darwin Core is intended for use in natural history collections, observation databases, and data repositories. EML is intended for use with data catalogs and data archives. CIF is used by the Cambridge Structural Database (CSD), described as a repository of crystal structures, whereas mmCIF is used to describe structures in the Protein Data Bank (PDB), another repository of structures. ThermoML is used to describe data in the ThermoML Archives as DDI is used to describe data sets stored in social science data archives. In each of these cases, the use of the terms database, data bank, catalog, archive, and repository reflect discipline-specific practices and different historical origins. There is nothing to suggest that archiving ThermoML property data is equivalent to archiving social science data sets, or that a repository of Darwin Core data is equivalent to a repository of CIF structures. Each of these comes from different traditions of data management. For example, social science archives are more closely associated with traditional library and digital archives (Blank & Rasmussen, 2004), whereas the ThermoML Archives are more closely associated with the OData (Open data protocol) and the World Data System (WDS). These different traditions, as well as the repurposing of schemes over time, make it difficult to understand the specific contexts for the goals articulated by individual communities.

Although many of the goals discussed above are independent of the discipline and type of data described, they are constrained by disciplinary perspectives and practices.

TABLE 7. Eleven fundamental requirements for metadata schemes for the documentation of scientific data.

Scheme abstraction	A well-defined metadata scheme will likely outlive its initial rendering. Abstraction allows needs be captured a way that supports multiple renderings over time.
Scheme extensibility, flexibility, and modularity	These are essential design requirements for information systems, including metadata, and will ensure the longevity of the scheme, facilitating adoption and modification over time and extension to meet yet-to-be-identified needs.
Comprehensiveness and sufficiency	Scheme creators should strive to define an element set (or vocabulary) that is comprehensive and also identify a minimal set of elements that are essential for documentation within the domain.
Simplicity	Scheme creators should take into account the levels of technical expertise of their community and support those with minimal as well as those with abundant tools and resources.
Data interchange (exchange)	An essential function of any metadata scheme for scientific data is the ability to exchange, share, and communicate data, whether raw or otherwise, among community members.
Data retrieval	Another essential function of any metadata scheme for scientific data is the ability to discover and acquire the data, taking into account discipline-specific access paths.
Data archiving	An essential function of any repository for the long-term preservation of information. Scheme developers should consider functions defined within the digital preservation community.
Data publication	Research data is an important component of the process of scientific communication and documentation. Scheme developers should account for the association of data with published research, such as citation in peer-reviewed journal articles.

In many cases, they capture general requirements for software systems, information systems, digital repositories, and communication, and address the differing levels of expertise within a community. Other goals, such as data comparability and validation, are also independent of discipline and type of data. Although these goals are not widely articulated, they likely reflect different practices across disciplines. Data comparability is expressed as a goal in both DDI and mmCIF; however, processes and requirements for comparison likely differ significantly across these disciplines. Data validation is a goal of CIF, ThermoML, and EML. Like comparability, the actual validation requirements likely differ across these disciplines. The only goal found to be specific to a domain or type of data is that of scheme harmonization, which was found to be more a goal of communities describing observational data. However, this too may be an artifact of community workflow and practice and not a difference limited by discipline or format. The topic of community workflow and practice is addressed in the section following the discussion of fundamental requirements for metadata for scientific data.

#### *Fundamental Requirements for Metadata for Scientific Data*

Requirements analysis is a necessary step of system design (Sommerville, 2005); and our review of the literature on discipline-specific schemes provides insight into community-driven activities in this area. For metadata to be more widely applicable across domains, more universal requirements gathering is needed. The review of similarities and differences conducted above provides a step in this direction, presenting a fair number of goals that are independent of the discipline and type of data described. The analysis also reveals goals that appear to be associated with specific types of data, but are applicable across disciplines.

In considering the different types of data, and the quest for making them accessible across domains and in an integrative manner, we have identified 11 goals that are foundational and universal, listed in Table 7. Seven of these goals—abstraction, extensibility, flexibility, modularity, comprehensiveness, sufficiency, and simplicity—are applicable beyond schemes for scientific data. The remaining four goals—data interchange, retrieval, archiving, and publication—are essential for any scheme proposing to support the use, reuse, and preservation of research data.

#### *Scope of Metadata for Scientific Data*

An initial question guiding this study addressed the scope of scientific data. The current analysis was limited to metadata schemes having sufficient accessible documentation and associated with publication on some level. This decision was informed, in part, by the authors' collaboration on the Dryad project (<http://www.datadryad.org>). As the work was pursued, it became apparent that scope was an important issue, and that another more fundamental question relating to scope may require examination. Although the study reported here was not suited to address our growing understanding of scope, we have considered this topic in the context of our research. As a result of this work, we propose that a significant question to study is how a community scopes their scientific data.

Scope is a broad term, but is commonly used in the software requirements and metadata communities to identify what is included as part of a system or scheme. In the context of metadata for scientific data, it seems that each community has scoped their metadata based on discipline-specific needs and practices. This observation makes sense, given that the metadata efforts examined are initiated within silos, embedded in the scientific practice of the community. To extend this research, it seems that more

questions are needed to address these fundamental requirements in the context of communities' approaches to science and communication.

Research examining scope seems essential given recent funding agency policy changes and developments within the digital archives and preservation communities (CCSDS, 2009; Higgins, 2008; National Institutes of Health, 2003; National Science Foundation [NSF], 2011a). The recent policy changes for preservation and access to research data are motivated by a common desire to change the conduct of scientific research (NSF, 2011b), largely driven by the broad adoption of computational techniques across disciplines. Standards emerging from the digital archives community are addressing the need for a common understanding of the purpose of archives and repositories. The specialized data archives and repositories responsible for the schemes studied here have demonstrated the accelerating effect of open access to research data on the scientific process. These exemplary systems (e.g., Interuniversity Consortium for Political and Social Research [ICPSR], PDB, CSD, and Gene Expression Omnibus [GEO]) have contributed to the transformation of their associated disciplines by providing centralized access to systematically organized data with a discipline-specific focus, serving as an active platform for secondary analysis and synthesis. Discipline-specific metadata schemes have improved the quality of documentation and facilitated alternate paths of access for researchers. Perhaps most important, highly specialized descriptive formats have permitted the provision of tools and services, including repositories.

The benefits noted directly above have been vital for scientific progress; however, they also contribute to artificial boundaries between disciplines and impede interdisciplinary and transdisciplinary reuse. As this study has demonstrated, many of the goals expressed by communities are independent of the discipline or type of data described. As new initiatives and governmental policies seek to change the conduct of scientific research, future efforts to define metadata schemes or vocabularies should adopt a more universal approach and look across disciplines to consider the lessons that can be learned from practices in other communities. Future research needs to examine in detail the historical and current practices across communities to better understand the workflow that underlies many of these goals and requirements. Research in this area needs to also study the technological context in which these schemes and systems were created. These steps are necessary to develop an understanding of requirements that are driven by technological trends and concerns. Future research should also examine in greater detail the specific functions supported by metadata schemes for scientific data, similar to earlier analyses of metadata for the organization and description of images (Greenberg, 2001). Finally, further research along the lines of that being conducted by Baker, Ribes, Millerand, and Bowker (2005) should be undertaken to better understand the role of metadata and standards in the larger and more complex picture of human organization in scientific communication and collaboration.

### *Limitations of This Study*

This study used a mixed-methods content analysis to examine the domains, objectives, and architectures of nine metadata schemes for documenting scientific data in the physical, life, and social sciences. The study used Greenberg's MODAL framework and categories derived from textual content describing each metadata scheme. The two key data sources for this are the selected metadata schemes, including their structure and properties, and textual documentation articulating the goals and intended function for the selected schemes. As with any study, the underlying methods, the sample (reliance on scheme documentation and schemes examined), and the guiding framework pose limitations.

The expanse of metadata systems developed for scientific data is extensive, and we limited our work to those areas with sufficient accessible documentation and with some connection to publication. Metadata schemes are complex systems, and the analysis of goals and functions underlying these schemes was intensive and time consuming. There may be limitations associated with scheme documentation and articles written conveying goals and intentions to potential adopters. We note this potential issue; however, the sufficient consistency among these sources of materials allowed us to conclude these were useful sources for this immediate study. We recognize that there are other schemes to consider and that further analysis might impact our findings. We also are aware that an additional analysis of metadata records might prove useful, but this activity is beyond this scope of this initial study. Despite these noted limitations, the work conducted is fairly extensive, and we believe the overall cohesiveness among the studied schemes provided solid base for pursuing the research reported on in this article. The limitations acknowledged point to future research studies that require scoping and will yield conclusions that would be useful to compare the findings presented in this article.

It should also be pointed out that the MODAL framework that guided our analysis is, in many ways, in a nascent state. In the work introducing MODAL, Greenberg (2005) states that "frameworks are artificial creations, with accompanying shortcomings," and that MODAL may "require enhancement and modification over time." She concludes by calling for testing of its applicability. Although MODAL proved to be the best suited framework for this research, we recognize that a different framework may yield different results. Despite these limitations, the study ran smoothly. In closing, the results provide insight into a topic of growing importance. The methods and procedures provide a framework that can guide further research in this area.

### **Conclusions and Future Direction**

The analysis presented in this article addresses the need for metadata-related research to inform more widely applicable metadata goals and requirements in support of data

sharing across disciplines and domains. Our research was conducted as a first step in this direction to better understand the scope and expanse of metadata systems that already exist. We were guided by a series of questions that helped with the examination of the similarities, differences, and scope of metadata schemes for the description of scientific data. A mixed-methods content analysis was conducted and Greenberg's MODAL framework was used to examine the principles, objectives, domains, and complexity of nine metadata schemes currently used in scientific data repositories. The article reviewed different types of scientific data, metadata-driven goals, research approaches, and frameworks for studying metadata, as well as the research methods and procedures applied in this study.

The nine metadata schemes analyzed in this study are representative of the physical, life, and social sciences as well as the three broad classes of experimental, observational, and statistical data. They are all used to document scientific data in repositories and span a period of several decades. The results presented in this study indicate that many of the goals expressed by communities during the scheme creation process are independent of scientific discipline or the type of data described. Many of these goals represent common requirements for software and system architectures or for digital archives and repositories without respect to domain. In addition to these findings, this study presented 11 fundamental goals or requirements for metadata schemes for the documentation of scientific data.

Through qualitative content analysis, 22 categories of metadata objectives were derived from textual content describing each metadata scheme. Relationships were identified between these objectives and the scheme domains (e.g., scientific discipline and type of data). For each strong correlation ( $>0.6$ ), a Fisher's exact test for non-parametric data was used to determine significance ( $p < .05$ ).

Significant relationships were found between the domains and objectives of the schemes. Schemes describing observational data are more likely to have "scheme harmonization" (compatibility and interoperability with related schemes) as an objective; schemes with the objective "abstraction" (a conceptual model exists separate from the technical implementation) also have the objective "sufficiency" (the scheme defines a minimal amount of information to meet the needs of the community); and schemes with the objective "data publication" do not have the objective "element refinement." These results support our conclusion that many of the goals driving metadata scheme creation are independent of the discipline or type of scientific data being described.

Research into the scope of metadata schemes seems essential given recent funding agency policy changes. Future research should examine in greater detail the context underlying these goals, including the community-specific practices and workflows as well as constraints caused by the technological environment and trends at the time of scheme creation. The research presented in this article is a step in

this direction. As noted in the Limitations section above, future research ought to extend to other schemes, and should also examine how metadata records adhere to or reflect scheme goals. There is also the need to explore frameworks focusing on metadata quality (e.g., Bruce & Hillman, 2004; Margaritopoulos et al., 2008; Stvilia et al., 2007), and how they may integrate and expand the work presented in this article. The results presented here provide an analysis of metadata schemes that has not been studied before, and help to break down the artificial barriers caused by metadata schemes created in silos. Moreover, the documented approach for study can be replicated or modified as this topic is pursued. These are important outcomes of the work presented in this article. In closing, we believe that continuing research in this area will enable us to contribute to the common goal of changing the conduct of scientific research through increased access to and reuse of research data.

## References

- Bain, J.L., & Michener, W.K. (2002). Ecological Archives: ESA's electronic data archive. *Ecological Society of America Annual Meeting Abstracts*, 87, 315.
- Baker, K.S., Ribes, D., Millerand, F., & Bowker, G.C. (2005). Interoperability strategies for scientific cyberinfrastructure: Research and practice. *Proceedings of the American Society for Information Science and Technology*, 42 (1). doi:10.1002/meet.14504201237
- Blank, G., & Rasmussen, K.B. (2004). The Data Documentation Initiative: The value and significance of a worldwide standard. *Social Science Computer Review*, 22 (3), 307–318.
- Bountouri, L., & Gergatsoulis, M. (2009). Interoperability Between Archival and Bibliographic Metadata: An EAD to MODS Crosswalk. *Journal of Library Metadata*, 9 (1–2), 98–133.
- Bosch, J. (2000). *Design and use of software architectures*. New York: Addison Wesley.
- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., . . . Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, 29(4), 365–371.
- Bruce, T.R., & Hillman, D.I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D.I. Hillman & E.L. Westbrook (Eds.), *Metadata in practice* (pp. 238–256). Chicago, IL: ALA Editions.
- Buschmann, F., Meunier, R., Rohnert, H., Sommerland, P., & Stal, M. (1996). *Pattern-oriented software architecture volume 1: A system of patterns*. New York: Wiley.
- Carrier, S., Dube, J., & Greenberg, J. (2007). The DRIADE Project: Phased application profile development in support of open science. In DC-2007: Application profiles: Theory and practice. International Conference on Dublin Core and Metadata Applications, Singapore, August 27-31, 2007, pp. 35–42.
- Chan, L.M., & Zeng, M.L. (2006). Metadata interoperability and standardization—A study of methodology, part 1. *D-Lib Magazine*, 12, 6.
- Committee on Science, Engineering, and Public Policy (US), and Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. (2009). *Ensuring the integrity, accessibility, and stewardship of research data in the digital age*. Washington, DC: National Academies Press.
- Consultative Committee for Space Data Systems (CCSDS). (2009). Reference model for an open archival information system—Draft recommended standard (CCSDS 650.0-P-1.1). Washington, DC: CCSDS Secretariat, National Aeronautics and Space Administration.

- Day, M. (1999). Metadata for digital preservation: An update. *Ariadne*, 22. Retrieved from <http://www.ariadne.ac.uk/issue22/metadata/>
- DCC/JISC. (2008). Infrastructure planning and data curation: A comparative study of international approaches to enabling the sharing of research data (ver. 1.6). Prepared by Raivo Ruusalepp Estonian Business Archives Consultancy.
- Duval, E., Hodgins, W., Sutton, S. & Weibel, S.L. (2002). Metadata principles and practicalities. *D-Lib Magazine*, 8(4).
- Frenkel, M., Chirico, R.D., Diky, V., Dong, Q., Marsh, K.N., Dymond, J.H., . . . Goodwin, A.R. (2006). XML-based IUPAC standard for experimentally predicted, and critically evaluated thermodynamic property data storage and capture (ThermoML). *Pure and Applied Chemistry*, 78(3), 541–612.
- Garvey, W.D. (1979). *Communication: The essence of science*. New York: Pergamon Press.
- Global Biodiversity Information Facility (GBIF). (2010). Darwin Core Archives. Retrieved from <http://www.gbif.org/informatics/standards-and-tools/publishing-data/data-standards/darwin-core-archives/>
- Greenberg, J. (2001). A quantitative categorical analysis of metadata elements in image applicable metadata schemas. *Journal of the American Society for Information Science and Technology*, 52, 917–914.
- Greenberg, J. (2003). Metadata and the World Wide Web. *Encyclopedia of Library and Information Science* (pp. 1876–1888). New York: Marcel Dekker.
- Greenberg, J. (2005). Understanding metadata and metadata schemes. *Cataloging & Classification Quarterly*, 40(3/4), 17–36.
- Greenberg, J. (2009). Metadata and digital information. *Encyclopedia of Library and Information Science* (pp. 3610–3623). New York: Marcel Dekker.
- Hall, S.R., Allen, F.H., & Brown, D. (1991). The crystallographic information file (CIF): A new standard archive file for crystallography. *Acta Crystallographica*, A47, 655–685.
- Hall, S.R., & McMahon, B. (2006). Genesis of the Crystallographic Information File. In S.R. Hall & B. McMahon (Eds.), *International tables for crystallography* (Vol. G, pp. 2–10). Hoboken, NJ: Wiley.
- Heidorn, P.B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, 57(2).
- Hey, A.J.G., & Trefethen, A.E. (2003). The data deluge: An e-science perspective. In F. Berman, G. Fox, & A.J.G. Hey (Eds.), *Grid computing—making the global infrastructure a reality* (pp. 809–824). Hoboken, NJ: Wiley.
- Higgins, S. (2008). The DCC curation lifecycle model. *The International Journal of Digital Curation*, 3(1), 134–140.
- Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain analysis. *Journal of the American Society for Information Science*, 46, 400–425.
- Hubenthal, U. (1998). Interdisciplinary thought. In W.H. Newell (Ed.), *Interdisciplinarity: Essays from the literature* (pp. 427–444). New York, NY: College Entrance Examination Board.
- Jones, M.B., Berkley, C., Bojilova, J., & Schildhauer, M. (2001). Managing scientific metadata. *IEEE Internet Computing*, 5(5), 59–68.
- Kelling, S. (2008). *Significance of organism observations: Data discovery and access in biodiversity research*. Copenhagen, Denmark: Global Biodiversity Information Facility.
- Klein, J. (1999). *Mapping interdisciplinary studies*. Washington, DC: Association of American Colleges and Universities.
- Kotani, M. (Ed.). (1975). Study on the problems of accessibility and dissemination of data for science and technology. *CODATA Bulletin* 16, 1–31.
- Krippendorff, K. (1980). *Content analysis: an introduction to its methodology*. Beverly Hills: Sage Publications.
- Lide, D.R. (1981). Critical data for critical needs. *Science*, 19, 1343–1349.
- Lim, S., & Liew, C.L. (2011). Metadata quality and interoperability of GLAM digital images. *Aslib Proceedings*, 63(5), 484–498.
- Margaritopoulos, T., Margaritopoulos, M., Mavridis, I., & Manitsaris, A. (2008). A conceptual framework for metadata quality assessment. In DCMI '08: Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications. Singapore: Dublin Core Metadata Initiative.
- Michener, W.K. (2006). Meta-information concepts for ecological data management. *Ecological Informatics*, 1(1), 3–7.
- Murray-Rust, P., & Rzepa, H.S. (1999). Chemical markup, XML, and the World Wide Web. 1. Basic principles. *Journal of Chemical Information and Computer Sciences*, 39(6), 928–942.
- National Institutes of Health. (2003). NIH data sharing policy. Retrieved from [http://grants2.nih.gov/grants/policy/data\\_sharing/](http://grants2.nih.gov/grants/policy/data_sharing/)
- National Science Board. (2005). Long-lived digital data collections: Enabling research and education in the 21st century. Retrieved from <http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf>
- National Science Foundation. (2011a). NSF data sharing policy. Retrieved from <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>
- National Science Foundation. (2011b). Changing the conduct of science in the information age. Arlington, VA: Author. Retrieved from <http://www.nsf.gov/pubs/2011/oi11003/oi11003.pdf>
- NSF Task Force on Cyberlearning. (2008). *Fostering learning in the networked world: The cyberlearning opportunity and challenge: A 21st century agenda for the National Science Foundation Report of the NSF Task Force on Cyberlearning*. Arlington, VA: National Science Foundation.
- Rayner, T.F., Rocca-Serra, P., Spellman, P.T., Causton, H.C., Farne, A., Holloway, E., . . . Brazma, A. (2006). A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, 7, 489.
- Ryssevick, J., & Musgrave, S. (2001). The social science dream machine: Resource discovery, analysis, and delivery on the web. *Social Science Computer Review*, 19(2), 163–174.
- Sommerville, I. (2005). *Software engineering*. Reading, MA: Addison-Wesley.
- Spellman, P.T., Miller, M., Stewart J., Troup C, Sarcines U, Chervitz S, . . . Brazma A. (2002). Design and implementation of a microarray gene expression markup language (MAGE-ML). *Genome Biology*, 3, 1–9.
- Spurgin, K.M., & Wildemuth, B. (2009). Content analysis. In B. Wildemuth (Ed.), *Applications of social science research methods to questions in library and information science*. Englewood, CO: Libraries Unlimited.
- Stvilia, B., Gasser, L., Twidale M.B., & Smith L.C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12), 1720–1733.
- Taxonomic Data Working Group (TDWG). (2009). Darwin Core. Retrieved from <http://rs.tdwg.org/dwc/>
- University of Kansas. (2004). The Darwin Core Version 1.3. Retrieved from [http://web.archive.org/web/20040405212931/http://tsadev.speciesanalyst.net/DarwinCore/darwin\\_core.asp](http://web.archive.org/web/20040405212931/http://tsadev.speciesanalyst.net/DarwinCore/darwin_core.asp)
- Westbrook, J.D., & Bourne, P.E. (2000). Star/mmCIF: An extensive ontology for macromolecular structure and beyond. *Bioinformatics*, 16(2), 159–168.
- Westbrook, J.H., & Grattidge, W. (Eds.). (1991). *A glossary of terms relating to data, data capture, data manipulation, and databases*. CODATA Bulletin, 23(1–2), 1–196.
- Zhang, Y., & Wildemuth, B. (2009). Qualitative analysis of content. In B. Wildemuth (Ed.), *Applications of social science research methods to questions in library and information science*. Englewood, CO: Libraries Unlimited.

## Appendix

### *Categories of Scheme Objectives Derived Using Content Analysis*

Category	Definition
Inter-scheme modularity	Elements from the scheme are intended to be used in conjunction with elements from other schemes (Duval et al., 2002) to meet new purposes.
Core set	The scheme is intended to provide only a core vocabulary, a common set of elements used to describe the most common situations. Darwin Core provides a “well-defined core vocabulary.” Schemes that have an objective of “comprehensiveness” generally do not have this objective.
Data lifecycle	The scheme is intended to support documentation of the data lifecycle—changes that occur to the data set over time.
Data portability	Data created using the scheme is intended to be “portable”—software application and operating system independent. This is generally an objective of older schemes.
Scheme simplicity	The scheme is intended to be simple and easy to use.
Data comparability	The scheme is intended to facilitate comparison of data sets.
Scheme stability	Conceptual stability—concepts represented in the scheme are stable and will not change over time (CIF, Darwin Core). Technical stability—the scheme implementation will not change, will be supported over time, and is safe to adopt.
Provenance	The scheme is intended to document the origin of information. This includes the origin of the data set (e.g., EML, DDI) or the origin of elements in the data set (e.g., ThermoML, mmCIF).
Element refinement	An aspect of scheme extensibility, element refinement is the ability to make more specific the meaning of an element (Duval et al., 2004). This is achieved through type extension (subclassing, deriving, subtyping). Refined elements can still be used in standards-based systems.
Scheme harmonization	The scheme is intended to be compatible and interoperable with other related schemes (e.g., DDI/ISO-11179, EML/GML) or the scheme was derived from an existing scheme (e.g., Darwin Core/Dublin Core, mmCIF/PDB).
Sufficiency (minimal set)	The scheme defines the minimal amount of information needed to achieve a specific goal for the community, for example, secondary data reuse (e.g., DDI, EML), experiment verification/reproduction (e.g., MAGE, MINiML).
Data validation	The scheme is intended to facilitate validation of data through the use of strongly typed data values.
Intra-scheme modularity	The scheme itself is modular and intended to support use of subsets of elements (or modules) for a particular purpose or particular stage of metadata creation. Modularity may also mean that data can be stored in multiple files (MAGE) or assembled at different times (DDI).
Abstraction	A conceptual model has been defined and is intended to be separate from the particular technical implementation.
Scheme flexibility	The scheme is intended to be adapted for use in settings outside of the current context.
Comprehensiveness	The scheme is intended to provide a comprehensive set of elements (or vocabulary) to describe a particular aspect of the domain. This is generally indicated by phrases such as “cover all” or “encompass all.” For example, ThermoML is intended to “cover all experimentally determined thermodynamic and transport property data.” The NEXUS scheme is intended to “encompass all information a systematist or phylogenetic biologic might wish to use.” CIF is a “precisely defined. . .comprehensive dictionary.”
Data archiving	The scheme is intended to facilitate the preservation/archiving of data sets and data documentation.
Data publication	The scheme is intended to support publication of data in journals and databases.
Data interchange	The scheme is intended to facilitate data interchange among community members—also referred to as data exchange, data sharing, or data communication.
Data retrieval	The scheme is intended to facilitate the discovery and acquisition of data.
Data documentation	The scheme is intended to describe not only the data, but to document the data context (experimental or observational context, analytical methods, etc.).
Scheme extensibility	The scheme is intended to be extended through the addition of new elements or modules to support future scheme growth or subdiscipline needs. When the scheme is extended, existing files or applications do not need to be modified. Extended files can still be used in standards-based systems.

*Classification of data from Kotani (1975)*

Categories of data	Description	Examples
Time-independent	Data that can be measured repeatedly	Most data in chemistry and physics, geological structures
Time-dependent	Data that can be measured only once	Volcanic eruptions, rare specimens, fossils
Location-independent	Data independent of location of objects measured	Most data in chemistry and physics, minerals, most data in biological sciences
Location-dependent	Data dependent on location of objects measured	Rocks, fossils, astronomical data, meteorological data, rare specimens, fossils
Primary observational or experimental data	Data obtained by experiment or observation	Optical spectra, crystallographic <i>F</i> values, seismographic records, weather charts
Derived data	Data derived by combining several primary data with the aid of a theoretical model	Fundamental constants, crystal structures, temperature distribution
Theoretical data	Data derived by theoretical calculations	Predicted solar eclipses
Determinable data	Data on a quantity which can be assumed to take a definite value under a given condition	Most macroscopic data
Stochastic data	Data on a quantity that takes fluctuating values from one sample or measurement to another	Polymer data, structure-sensitive properties, soil composition, solar flares, most metrology
Quantitative data	Measures of scientific quantities in terms of well-defined units	Most data in chemistry and physics, seismic data, meteorological data
Semiquantitative data	Measures of scientific quantities using arbitrary scales	Mohs hardness scale, wind force scale
Qualitative data	Any scientific definitive statement concerning scientific objects	Chemical structure formulas, properties of nuclides, rock classification, amino-acid sequences
Data as numerical values	Data presented as isolated numerical values	Meteorological data
Data as models or graphs	Data presented in graphical form or as models	Phase diagrams, molecular models, geologic maps, genetic pathways
Symbolic data	Data presented using arbitrary symbols (non-numeric)	Lithology in borehole data