Geochemistry Geophysics Geosystems G3
AN ELECTRONIC JOURNAL OF THE EARTH SCIENCES
Published by AGU and the Geochemical Society

Technical Brief
Volume 4, Number 1
25 January 2003
1010, doi:10.1029/2002GC000318
ISSN: 1525-2027

Scalable models of data sharing in Earth sciences

John Helly
San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, Mail Code 0527, La Jolla, California 92093, USA (hellyj@ucsd.edu)

Hubert Staudigel and Anthony Koppers
Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California, San Diego, 8800 Biological Grade, La Jolla, California 92037, USA (hstaudigel@ucsd.edu; akoppers@ucsd.edu)

[1] Many Earth science disciplines are currently experiencing the emergence of new ways of data publication and the establishment of an information technology infrastructure for data archiving and exchange. Building on efforts to standardize data and metadata publication in geochemistry [*Staudigel et al.*, 2002], here we discuss options for data publication, archiving and exchange. All of these options have to be structured to meet some minimum requirements of scholarly publication, in particular reliability of archival, reproducibility and falsifiability. All data publication and archival methods should strive to produce databases that are fully interoperable and this requires an appropriate data and metadata interchange protocol. To accomplish the latter we propose a new Metadata Interchange Format (.mif) that can be used for more effective sharing of data and metadata across digital libraries, data archives, and research projects. This is not a proposal for a particular set of metadata parameters but rather of a methodology that will enable metadata parameter sets to be easily developed and interchanged between research organizations. Examples are provided for geochemical data as well as map images to illustrate the flexibility of the approach.

Components: 4927 words, 12 figures.

Keywords: Data management; publication; metadata; geosciences; interdisciplinary; data sharing.

Index Terms: 9810 General or Miscellaneous: New fields (not classifiable under other headings); 1094 Geochemistry: Instruments and techniques; 1594 Geomagnetism and Paleomagnetism: Instruments and techniques; 3094 Marine Geology and Geophysics: Instruments and techniques; 4594 Oceanography: Physical: Instruments and techniques.

Received 25 January 2002; **Revised** 7 August 2002; **Accepted** 15 August 2002; **Published** 25 January 2003.

Helly, J., H. Staudigel, and A. Koppers, Scalable models of data sharing in Earth sciences, *Geochem. Geophys. Geosyst.*, *4*(1), 1010, doi:10.1029/2002GC000318, 2003.

1. Introduction

[2] Most data and metadata in earth sciences are published in the context of traditional, peer-reviewed publications in paper journals. The practice of publishing data electronically is extremely poorly developed, in particular, since electronic journals continue to be functionally similar to paper journals. The confinement of most data publication to paper journals has the result that authors rarely publish data and even less frequently their metadata [*Helly*, 1998]. In geochemistry, this has resulted in a crisis in data publications (GERM Steering Committee, available at http://earthref.org/events/GERM/2001/lajolla-01-announcement.htm, 2001) where a large amount of legacy data are

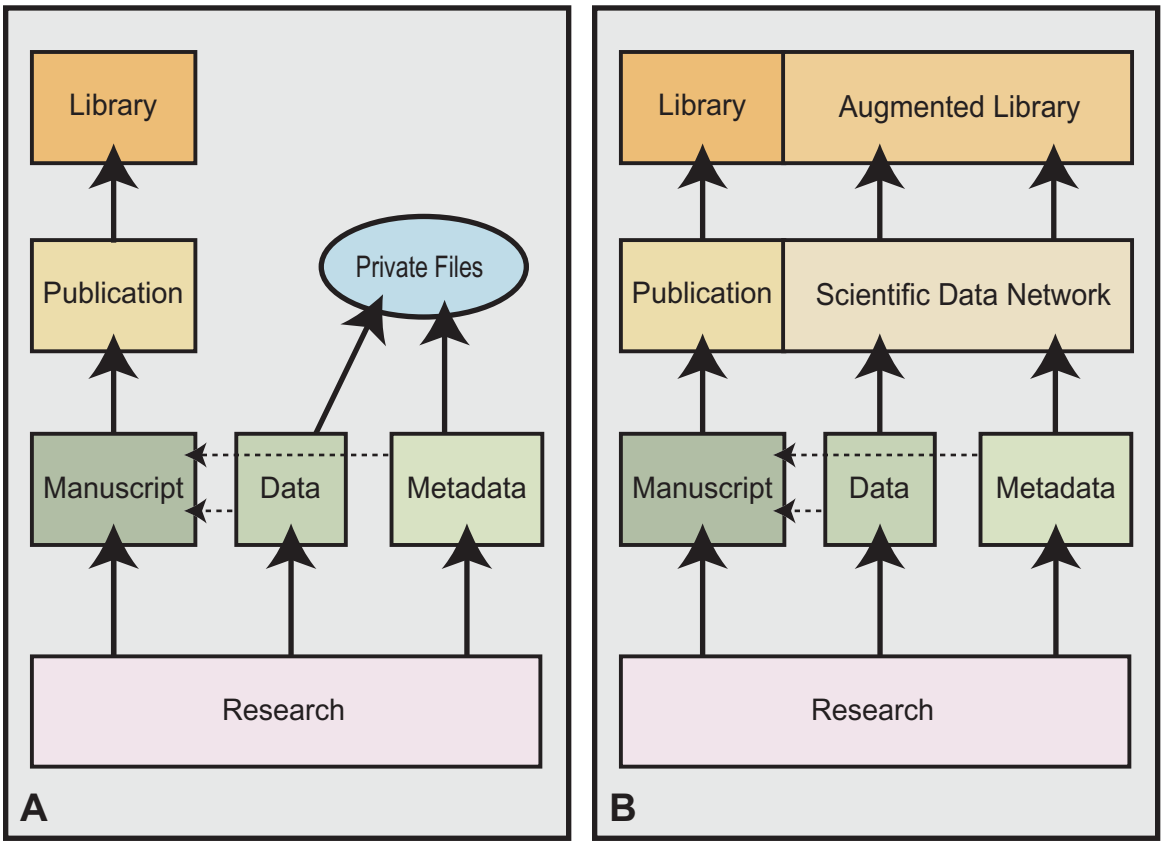


Figure 1. (a) Schematic depiction of the flow of scientific information from research to published library resources as currently practiced. (b) Potential approach based on interoperable resources contained within a scientific data network but leading to library-grade products analogous to the formally published journals we are familiar with today. This augmented library will protect the future of science against any loss of valuable research data that normally resides in private files. Note that the dashed lines from data and metadata to the manuscript reflect the limited publication of these information sources in our conventional manuscripts.

in danger of being permanently "buried" in private research files (see Figure 1a). earth sciences are likely to gain substantially from introducing new means of publishing data. There is a wide array of electronic data publications that range from strict peer-reviewed, central publication to entirely non-reviewed, decentralized publication on personal websites (Figure 2).

[3] Research produces effectively three types of data products: (1) data, (2) metadata and (3) the interpretations of these data in the form of various illustrations included in manuscripts. Our current paper publishing protocol (Figure 1a, dashed lines) is mostly biased toward the final writing and illustrations. Digital data and metadata tend to be rarely published on paper copy or in its electronic equivalents (dashed lines). Page limitations in high profile journals actively work to eliminate or greatly reduce the actual publication of data and metadata. As a result, interpretations and figures based on data are widely published and archived in libraries, while most of the primary data are confined to research files of investigators. These private archives, however, do not provide sufficient access for future research that might result in a reinterpretation of the data. There exists no return flow of data into community-wide research activities and, therefore, these data are effectively lost. In the worst possible case, samples will have to be reanalyzed, forcing another cycle of data generation and loss. This puts earth sciences in a situation where the transient interpretations of data are kept in reliable archives while the actual permanent measurements and records have an uncertain fate. This effective loss of data or metadata due to non-publication cannot be

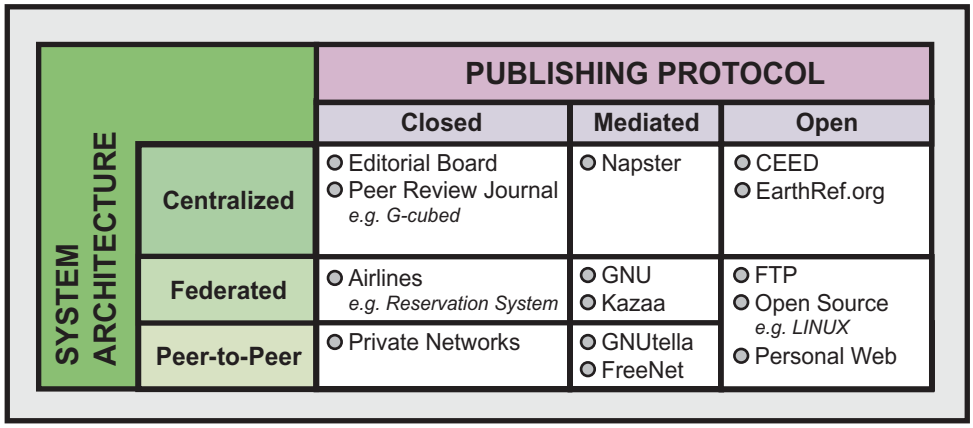


Figure 2. Trade-off between system architecture and publishing protocol providing possible models of data publishing and sharing. Note that this table is merely illustrative and not exhaustive.

reconciled with our principles of scholarly science. Without them, scientific work cannot be duplicated or falsified.

[4] In Figure 1b, we show an alternate data publication protocol in which the generation of scientific products and its publication remains the same, but data and metadata are published in a scientific data network. This network takes on the role of libraries for the archiving and serving of data to the scientific community in order to support future research. In this new protocol the traditional publication of scientific results is accompanied by publication of its data and metadata making the ''complete'' scientific product available in a consistent and coherent manner [*Baru et al.*, 1997; *Helly et al.*, 2002]. Such a data network may include a range of options, including the attachment of electronic data supplements to publications (GERM Steering Commmittee, available at http://earthref.org/events/GERM/2001/lajolla-01-announcement.htm, 2001).

[5] Electronic data publication in earth sciences is still in an emerging phase, where many options may be considered that have particular advantages to particular earth science communities. Metadata play an important role in such data networks, but as different earth science disciplines may choose different data publishing protocols, it becomes extremely important to have data and metadata in representations that are interchangeable between protocols, standards and conventions. These features are referred to as interoperability and platform-independence.

[6] In this paper we will discuss different methods of data publishing via a scientific data network, the desirable characteristics required to make such a network into a truly scholarly resource, and we will introduce a method for generating a highly portable metadata interchange format we call .mif (pronounced dot-mif) that will help make an earth science data network interoperable. This .mif format is designed to help search, extract and exchange data and metadata as they might be stored in a range of personal or community earth science archives. We will conclude this paper with a discussion of how this metadata format can be scaled to support the diversity of interests within earth science communities. This paper is part of a discussion on metadata and data archiving infrastructures, and the most recent contributions to this discussion may be found at http:earthref.org\metadata\GERM\.

2. Methods of Data Publishing

[7] Geochemical and earth science data may be published in a variety of ways. Information Technology (IT) offers a range of publication options that may be classified by their respective publishing protocols and their system architecture. In Figure 2 we have illustrated various publication protocols using these two parameters. The choices between these protocols are mainly driven by the technical issues relating to centralization, peer-review procedures, the complexity of data, the cost of publication and archival requirements. However, some of the most difficult choices often involve

sociopolitical issues, such as which groups or organizations are interested in operating a data network whether centralized or decentralized. These choices are difficult because they usually have to be made competitively where the winner is expected to deliver the best product for the community, in exchange for funding and scientific benefits. A careful evaluation is needed to balance the benefits of the freedom of managing one's own data versus the responsibilities and costs of maintaining a large number of public data servers that are also reliable.

[8] The classical model of scholarly publishing is the scientific journal with a protocol based on peer-review and editorial boards. The architecture supporting this is a centralized one typically located at the publisher's facilities where the copyright and copy-of-record are retained. In the past, libraries were also keepers of the copy-of-record and the redundancy of libraries provided insurance that there would "always" be a copy available. Now, with the shift to e-publishing (electronic publishing), it is the publishers who hold the copy-of-record, and the libraries themselves never actually hold anything. Subscriptions to electronic journals give temporary access to contents, not ownership of an archive. At the end of the spectrum of centralization is a completely decentralized system architecture united only by the basic http protocol is exemplified by individual web-sites without any publishing protocol per se; what you see is what you get and it may not be there tomorrow.

[9] As an intermediate case, federated networks may be thought of as decentralized but locally structured and operated according to shared, community-defined protocols and standards. A particularly interesting and recent example is the music distribution network approach of Kazaa at http://www.kazaa.com. Kazaa uses a peer-to-peer approach but selects a subset of peers to also act as search-server nodes according to their exceptional computational and communication resources. This has the particular advantage of federating the server function that supports searching (based on the metadata of song, artist and peer-node) without centralizing it. This approach is also capable of dynamically reconstituting the subset of servers if one or more is removed from the network for any reason. Consequently, it is referred to as a self-organizing network.

[10] In the earth sciences, disciplinary boundaries will naturally define important nodes in a scientific data network with discipline-specific publishing protocols and conventions. It could be operated as a self-organizing network (such as Kazaa) given sufficient standardization of both the data and metadata formats [see also *Staudigel et al.*, 2002]. Later in this paper we describe a new approach to metadata standards that could serve as a further step toward such a self-organized network. These protocols and conventions must provide local structure for the operation of the scientific data network while supporting the diversity that is the very heart of specific earth science research. Such protocols and conventions would include metadata standards, conventions for naming arbitrary data objects (ADO) so they can be uniquely cited and versioned (to enable verification/falsification) and conventions for protecting intellectual property rights. However, to clarify this notion we consider next the features that a scientific data network should have in order to effectively support scientific research in the future. As earth Science disciplines evolve their data management approaches, many options are possible. Key to any approach, however, is a design that supports scholarly work, and an effective data exchange mechanism.

3. Desirable Characteristics of a Scientific Data Network

[11] Any scientific data network has to be designed to serve some basic principles of scholarly publication. One of the most fundamental requirements is the verification and falsification of results. Any analysis in science has to be accompanied by its data and metadata foundation during the process of the analysis, but also during its publication and archival. This dictum demands reliable access to published data and the ability to unambiguously identify the data used in any given analysis. To support these requirements, we have reproduced a table of essential functions needed in a data publication system (Figure 3). These functions should

FUNCTION	PURPOSE
User Registration	Assignment of an ***user ID*** and ***password*** to a given user while acquiring their email address and related contact information. This is used for auditing data access and communication with users.
Data Acquisition	Data contribution or submissions through ***uploading*** while acquiring a minimal set of ***metadata***. This initiates the automatic creation of an unique and persistent name for the ADO and a transportable metadata file which is directly bundled with the uploaded ADO.
Search	A search system providing for ***spatial***, ***temporal*** and ***thematic*** queries based on the content of the uploaded metadata.
Deletion Control	The ability to ***delete*** an ADO must be tightly controlled to prevent the arbitrary deletion of data that users already have downloaded. In a manner analogous to journal articles, one should ***not*** un-publish data. Errata can be accomodated by publishing a revision of the data as a ***new*** ADO version. A special case to consider is the editorial peer-review process. This requires confidentiality and the ability to remove an ADO if not accepted for peer-reviewed publication. A looser deletion policy would allow deletion of data if it had never been downloaded by an user.
Assignment of Persistent Names	ADO's within a data repository should have ***persistent*** names. This allows for monitoring when updates of ADO's come online. It helps identifying which ADO's are most frequently downloaded or may be used to notify users of anomalies/issues related to a particular ADO. Finally it establishes precedence by ***publication date*** and enables citability in publications.
QA/QC	***Quality assurance and control*** can exist to varying degrees. It is exemplified by peer-review or non-peer-review and by anomaly detection and reporting. It must be explicitly stated. Some investigation is beginning on approaches to ***semi-automate*** this function for specific types of data.
Access Control	Access control enables the ***data contributor*** to specify a password known only to him/herself for an uploaded ADO. This password may subsequently be provided to other users to assure limited access. This enables data submitters to ***independently control*** access to their published data. Any user attempting to retrieve a password-protected ADO from the system must first obtain that password from the contributor of the data.
Versioning Traceability	Versioning data is required in order to prevent ***in situ*** changes to data. This enables the establishment of ***data heritage*** which is required to inform the user of the emperical, derived or computed ***nature*** of the data. This is neccesary to ensure the reproducibility of results and it reserves intellectual property rights and facilitates proper attribution.

Figure 3. Basic functions for controlled publication of scientific data after *Helly et al.* [2002]. ADO refers to an Arbitrary Digital Object which is a general term to describe any digital object, in general a file of some kind, that can be stored on a computer system.

be explicitly addressed in any system architecture in order to preserve all scientific data and metadata, regardless of the degree of its centralization.

[12] In addition to these functions, we must consider reliability and availability of such a data network to ensure ready and repeatable access. These system engineering concerns are frequently overlooked in discussions about scientific publishing whether for data or scientific manuscripts. Reliability and availability of these electronic resources are vital with respect to the conduct of science and the equitable access to data throughout the research community. The software community has achieved this, albeit imperfectly, by a high degree of redundancy through the use of mirror sites for software, and a painstaking

emphasis on clearly articulating the software version and hardware dependencies directly in the name of a software distribution file, for example. Similar efforts of redundancy may be a key to the longevity of earth science data and metadata archival and effective data exchange between interoperable databases is essential.

4. Designing Metadata for Data Sharing

[13] Metadata are essential to a successful data sharing, but the term can be very confusing. Metadata is many things to many people [*Baru et al.*, 1997; *Daniel and Lagoze*, 1998; Federal Geographic Data Committee, available at http://www.fgdc.gov/metadata/contstan.html, 1998; PURL, Dublin Core Metadata Initiative, available at http://dublincore.org/index.shtml, 2002; S. Weibel, Dublin Core Metadata, available at http://purl.org/metadata/dublin_core, 1998]. Our focus here is on key functions of metadata that are relevant to research and data exchange in the earth sciences. Two main types of metadata may be distinguished. First, metadata is used for discovering the existence of data by searching a metadata catalogue or its equivalent. Second, metadata is documentary information describing the content, context, quality, structure, accessibility and so on of a specific data set.

[14] In this latter respect metadata plays a role in the integration of data, analysis and modeling. In this paper we, therefore, refer to metadata either as cataloguing metadata (i.e., used for searching and archiving) or application metadata (i.e., used for data integration and scientific analysis). The key to these distinctions between data and metadata is how the information is used. Frequently, any particular parameter may be used in either way depending on the user's purpose. For example the geographic latitude may be considered as data in the context of a correlation of rain isotope data with latitude, or it may be considered to be catalogue metadata when it is used to populate and search a metadata catalogue for papers relevant to particular regions. In another context, it may be considered application metadata when it is used as a variable in data analysis. Application metadata describing analytical methods (such as Appendix 2a–2d, in *Staudigel et al.* [2002]) may also be treated as data when they are shared between databases. Once shared, these same data may be used as both catalogue metadata to locate samples analyzed by a particular method as well as application metadata to support a data integration process merging samples analyzed by different methods and requiring a conversion of units.

[15] Consequently, discussions on metadata can be confusing, subjective and discouraging. To make any progress on this front one seeks a systematic way to approach the definition of a reasonable set of metadata parameters. We begin this by recognizing that metadata is fundamentally arbitrary in nature. What's important to one individual may be of little consequence to another and who's to say what is more important? However, there usually exist subsets of metadata that are clearly required for any scholarly publication and more or less effortlessly agreed upon and that can be augmented to accommodate more specialized metadata requirements as they emerge. So we have designed an approach that accommodates this essential arbitrariness but enables compliance with whatever metadata standards or conventions emerge within the scientific, information technology, library and publishing communities. Some of the diversity of efforts that are relevant to the earth science community, and that must be accommodated, is reflected in the list presented below:

- Federal Geographic Data Committee http://www.fgdc.gov/metadata/contstan.html
- United States Geological Survey http://www.usgs.gov/tools/metadata
- World-Wide Web Consortium (W3C) http://w3.org
- Earth Science Markup Language http://esml.itsc.uah.edu/products.html
- Canadian Geochemistry Online http://geochem.gsc.nrcan.gc.ca/
- Ecological Metadata Language http://knb.ecoinformatics.org/
- Object Management Group (OMG) http://www.omg.org
- Corporation for National Research Initiatives http://www.handle.net
- Dublin Core http://www.dublincore.org

[16] Each of these approaches has strengths and weaknesses. For example, the Federal Geographic Data Committee (FGDC) standard was designed for geographic and remote sensing data and not for survey-type data, or laboratory data. It has subsequently been amended to a certain degree to make these accommodations and this emphasized the need to have a flexible convention that can evolve in response to disciplinary needs. It is clear is that there is a proliferation of metadata conventions that will continue to grow to accommodate particular purposes. It is, therefore, important to develop methods of interchanging metadata across these various conventions with ease and reliability to enable digital collections of data and digital libraries to interoperate making data more widely accessible.

[17] In the design of metadata it is helpful to think in terms of data structures with well-understood mathematical and computational properties. One such data structure is a tree: a hierarchy with a root node, branches, subtrees and leaves. Given categories of metadata, it is in some sense natural to organize them into a hierarchy or tree (Figure 4). Other data structures are possible and each has its own properties. However, a tree provides a modular structure that enables the addition and deletion of subtrees of arbitrary depth and breadth without affecting other parts of the tree and this is well-suited to an evolving metadata architecture. For example, some sub-trees can be used for several earth science disciplines without any adjustments because of commonalities in information requirements; such as the geographic data for sample location. A hierarchical structure also allows for the inclusion of other, well established metadata formats, such as the Dublin Core for bibliographic information that is used in almost any scientific effort (Figure 4). Each metadata parameter is a leaf attached to a branch of the tree and a path from the root to any leaf provides a complete and unique description of that metadata parameter. The same parameter name, for example latitude, may be used an arbitrary number of times as a leaf throughout the tree while each leaf is uniquely identifiable by the path through the tree to that particular use of latitude.

[18] In Figure 4 we show a tree rooted to a unique identifier for the metadata interchange format (.mif) and including sub-trees for applications and catalogues (e.g., water analyses and the Dublin Core metadata standard). As we will show below this type of tree structure can be cast into a form that can be conveniently represented in a flat, ASCII-encoded, computer file.

5. Metadata Interchange Format

[19] The main goal for our metadata interchange format (*.mif) is to facilitate the automated electronic extraction and transfer of data and metadata from and between databases as well as its accurate and effective reuse. The .mif convention allows a program to recognize the structure and contents of an electronic file. In our examples we use the spreadsheet compilations of data and metadata as they are described in the companion paper by *Staudigel et al.* [2002]. These files are considered ADOs and the *.mif files describe these files in a way that allows a database to export its contents and to identify its contents and structure, so the tables can be uniquely imported (i.e., read, accurately reproduced and translated into other formats) by another data system. Such an export/import approach makes a database interoperable with others by facilitating automated access to its contents.

[20] In Figure 6 we present an example of the *.mif format for the ROCK SAMPLE MAJOR ELEMENTS table from *Staudigel et al.* [2002]. The *.mif can be described as a series triples (i.e., three elements) that describe the headers of a data table. Each triplet consists of a narrative label, a name for the metadata parameter, and a value. The label field provides a place to put a conventional, human-readable label that may contain special characters that are not well-suited to inclusion in metadata catalogues as searchable metadata (e.g., some special characters are used as delimiters in operating system commands and this can result in interference with programs with abnormal results) but that have special significance in the context of the data object itself. The parameter name is constructed so that it encodes the logical structure of the ADO and enables it to be accurately reconstructed and interpreted

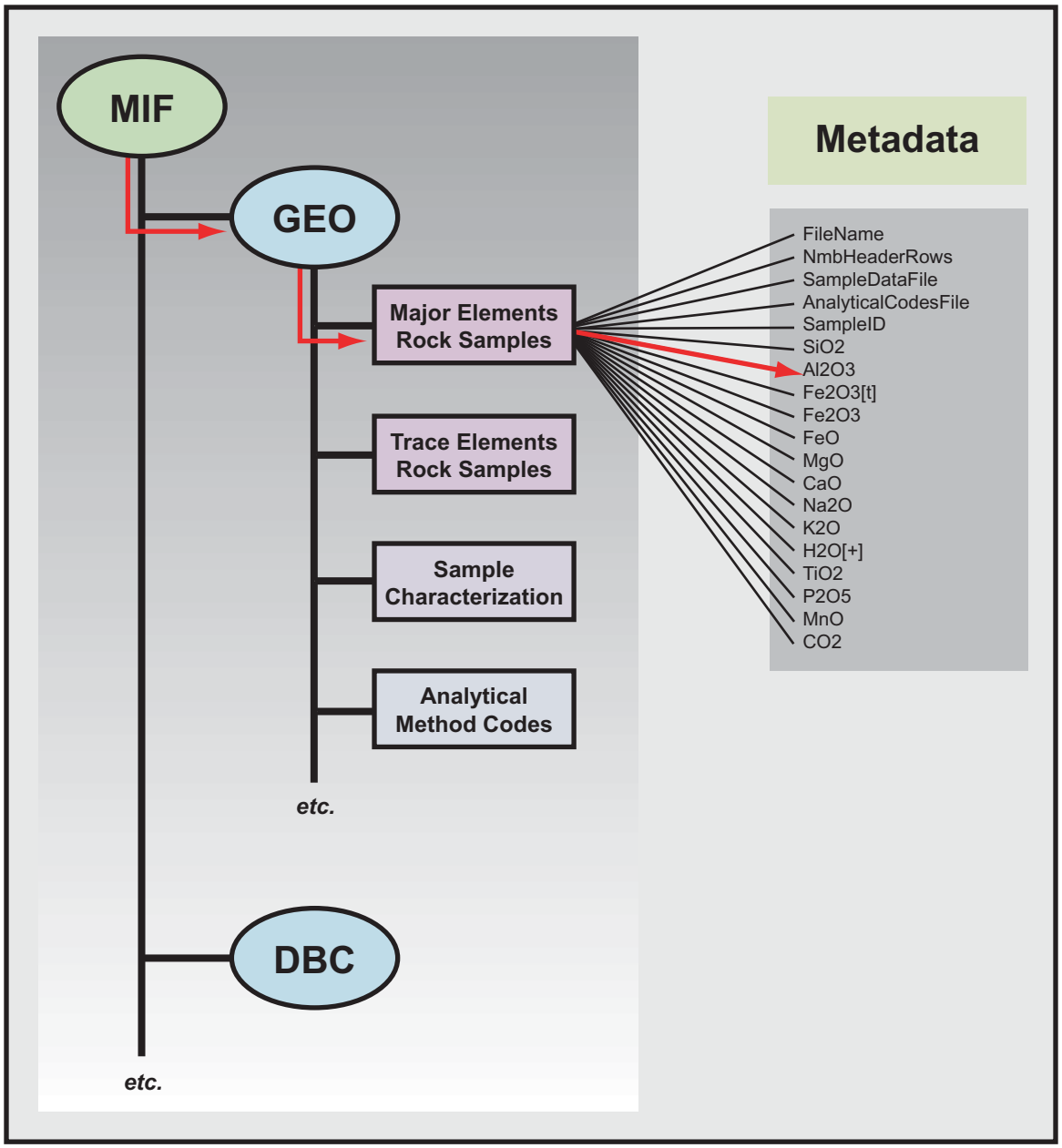


Figure 4. Example metadata tree. The root is at MIF and the leaves are at the right. Red indicates path corresponding to *.mif record shown in Figure 5. DBC corresponds to other non-GEO metadata content standards that might be also be supported by the community such as Dublin Core.

from the metadata content. The values field provides a set of metadata that can be stored in a catalogue to enable the ADO to be found and to provide cross-references to related ADOs containing information relevant to, in this case, the major element analyses as described in Appendix 3 of *Staudigel et al.* [2002]. For example, among the metadata fields in Figure 6 are also fields that specify the ADO containing the analytical code metadata (i.e., staudigel_analcodes.csv) and the positional information (i.e., staudigel_sampdata.csv). The latter information enables the major element table ADO to be discovered in a geospatially based search. Alternatively, the geospatial information could be included in the metadata in Figure 2. These are further examples of the use of data and metadata interchangeably as a function of the purpose they are used for. Each record of the corresponding comma-

parameter	SiO2	Al2O3	Fe2O3[t]	Fe2O3	FeO	MgO	CaO	Na2O	K2O	H2O[+]	TiO2	P2O5	MnO	CO2
unit	wt%	wt%	wt%	wt%	wt%	wt%	wt%	wt%	wt%	wt%	wt%	wt%	wt%	wt%
analytical code	1	2	3	4	5	6	7	8	9	10	11	12	13	14
CY-19.7a	48.8	16.56	7.99	6.97	0.92	4.54	6.99	1.39	4.72	4.77	0.51	0.04	0.09	2.39
CY-19.7b	47.42	16.39	8			4.64	6.93	1.59	4.68		0.52	0.03	0.08	
CY-19.7c	47.1	16.2	7.57			4.5	6.56	1.49	4.58		0.48	0.02	0.09	
CY-23.2a	45.9	14.6	8.15	7.01	1.03	3.73	7.32	0.6	7.55	4.68	0.57	0.08	0.06	5.03
CY-23.2b	45.5	14.6	9	8.78	0.2	3.62	7.52	0	7.62	4.4	0.57	0.1	0.06	6.4
CY-23.3a	44.9	13.7	7.87			3.94	7	0.59	7.32		0.52	0.13	0.06	
CY-32.3b	47.5	17.2	8.48			7.07	3.28	1.42	4.18		0.58	0.03	0.08	
CY-32.3c	49.5	18.03	8.77	7.65	1.01	7.03	3.51	1.39	4.24	6.25	0.69	0.04	0.08	0.24
CY-32.3d	47.7	17.6	8.6	7.49	1	6.78	3.64	1.3	4.24	6.9	0.68	0.05	0.09	0.4
CY-32.3e	48.53	17.75	8.75			6.87	3.48	1.12	4.25		0.7	0	0.08	

Figure 5. Major Element Table from *Staudigel et al.* [2002].

LABEL	METADATA PARAMETER	VALUE
MIF Version	MIF_Version	1.0.2
ADO	GEO_Major-Elements-Table_FileName	staudigel.rocks.csv
Header Rows	GEO_Major-Elements-Table_NmbHeaderRows	1
Sample Data	GEO_Major-Elements-Table_SampleDataFile	staudigel.sampledata.csv
Analytical Codes	GEO_Major-Elements-Table_AnalCodeFile	staudigel.analcodes.csv
Sample ID	GEO_Major-Elements-Table_SampleID	col=1
SiO2	GEO_Major-Elements-Table_SiO2	col=2:ac=1:unit=wt%
Al2O3	GEO_Major-Elements-Table_Al2O3	col=3:ac=2:unit=wt%
Fe2O3[t]	GEO_Major-Elements-Table_Fe2O3t	col=4:ac=3:unit=wt%
Fe2O3	GEO_Major-Elements-Table_Fe2O3	col=5:ac=4:unit=wt%
FeO	GEO_Major-Elements-Table_FeO	col=6:ac=5:unit=wt%
MgO	GEO_Major-Elements-Table_MgO	col=7:ac=6:unit=wt%
CaO	GEO_Major-Elements-Table_CaO	col=8:ac=7:unit=wt%
Na2O	GEO_Major-Elements-Table_Na2O	col=9:ac=8:unit=wt%
K2O	GEO_Major-Elements-Table_K2O	col=10:ac=9:unit=wt%
H2O[+]	GEO_Major-Elements-Table_H2Oplus	col=11:ac=10:unit=wt%
TiO2	GEO_Major-Elements-Table_TiO2	col=12:ac=11:unit=wt%
P2O5	GEO_Major-Elements-Table_P2O5	col=13:ac=12:unit=wt%
MnO	GEO_Major-Elements-Table_MnO	col=14:ac=13:unit=wt%
CO2	GEO_Major-Elements-Table_CO2	col=15:ac=14:unit=wt%

Figure 6. Example of the metadata interchange format (*.mif) using the rock sample major elements table from *Staudigel et al.* [2002].

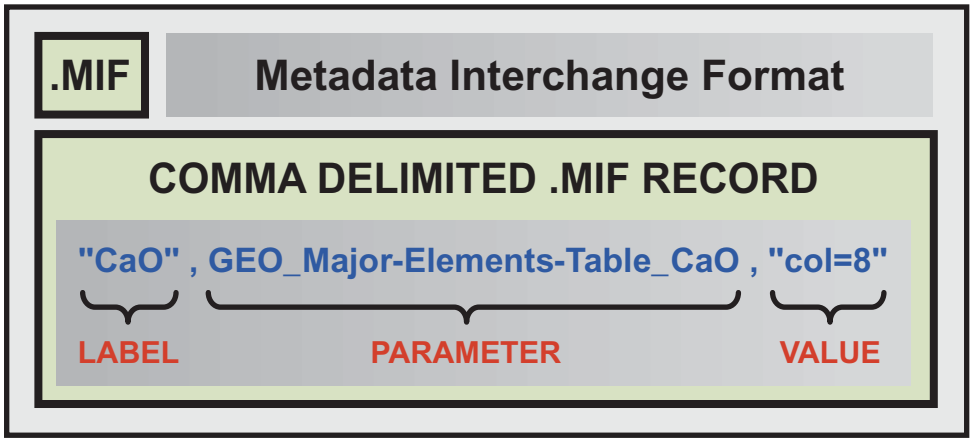


Figure 7. A single metadata record in *.mif. Underscores separate levels of the hierarchy in this example and hyphens delimit words that would otherwise be separated by blank spaces. These conventions are used to ensure proper parsing for import and export.

delimited, ASCII, row-oriented *.mif file would be formatted as shown in Figure 7.

5.1. Defining a Standard Set of Metadata Parameters

[21] The .mif standard may evolve with time, or different databases may employ different .mif structures. To allow for conversions between such versions or database structures, the *.mif convention includes a metadata template file (*.mtf) describing a particular version of the interchange metadata format standard. The relevant science community can control the content of the *.mtf file through, for example, a working group by defining a standard set of parameters that are used as metadata for the various standard ADOs. This set can be as large as necessary but as small as is sufficient. There should be at least one metadata group or block per type of ADO. In our rock-sample-major-elements-table example above, this type of table is defined and published as a community-recognized type of digital object (i.e., ADO), the content of its associated metadata file (e.g., staudigel_rocks.mif) is published in accordance with the definition of metadata parameters contained in a published versions of a *.mtf (e.g., AGU_GEO.mtf) and researchers can freely exchange this type of object with its standard metadata in a simple and unambiguous way while maintaining a common and consistent interface as illustrated in Figure 8.

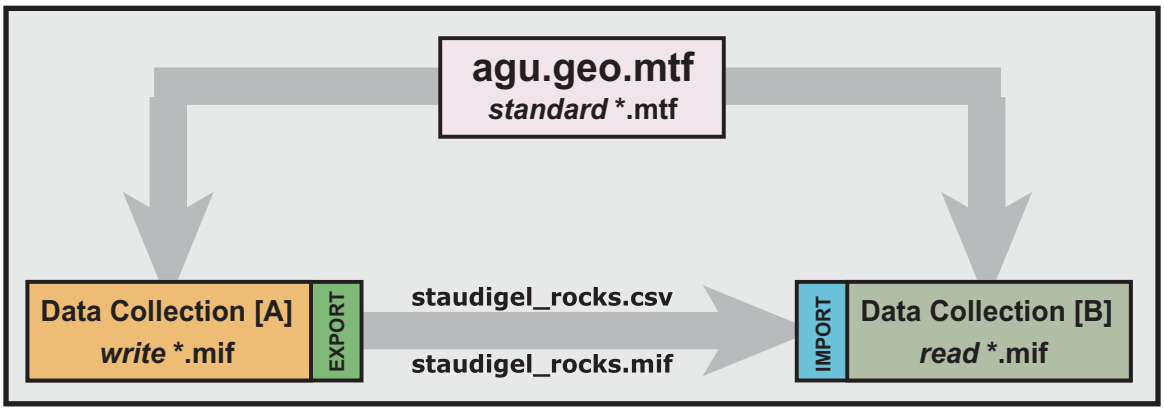


Figure 8. Depiction of the use of *.mtf files to define how the *.mif files should be written and read by the data exporter and data importer respectively. The *.mif file describes how to read and interpret the ADO (i.e., staudigel_rocks.csv). The *.mtf file contains a metadata block for the rock-sample-major-elements-table ADO and therefore defines how the *.mif file for the ADO should be written and therefore defines the structure of the ADO.

LABEL	METADATA PARAMETER	DEFAULT	DATA TYPE	
MIF Version	MIF_Version	1.0.2	character	40
ADO	GEO_Major-Elements-Table_FileName	<ADO Name>	character	128
Header Rows	GEO_Major-Elements-Table_NmbHeaderRows	3	integer	40
Sample Data	GEO_Major-Elements-Table_SampleDataFile	<ADO Name>	character	128
Analytical Codes	GEO_Major-Elements-Table_AnalCodeFile	<ADO Name>	character	128
Sample ID	GEO_Major-Elements-Table_SampleID	<col>	character	40
SiO2	GEO_Major-Elements-Table_SiO2	<col:ac:unit>	real	10.2
Al2O3	GEO_Major-Elements-Table_Al2O3	<col:ac:unit>	real	10.2
Fe2O3[t]	GEO_Major-Elements-Table_Fe2O3t	<col:ac:unit>	real	10.2
Fe2O3	GEO_Major-Elements-Table_Fe2O3	<col:ac:unit>	real	10.2
FeO	GEO_Major-Elements-Table_FeO	<col:ac:unit>	real	10.2
MgO	GEO_Major-Elements-Table_MgO	<col:ac:unit>	real	10.2
CaO	GEO_Major-Elements-Table_CaO	<col:ac:unit>	real	10.2
Na2O	GEO_Major-Elements-Table_Na2O	<col:ac:unit>	real	10.2
K2O	GEO_Major-Elements-Table_K2O	<col:ac:unit>	real	10.2
H2O[+]	GEO_Major-Elements-Table_H2Oplus	<col:ac:unit>	real	10.2
TiO2	GEO_Major-Elements-Table_TiO2	<col:ac:unit>	real	10.2
P2O5	GEO_Major-Elements-Table_P2O5	<col:ac:unit>	real	10.2
MnO	GEO_Major-Elements-Table_MnO	<col:ac:unit>	real	10.2
CO2	GEO_Major-Elements-Table_CO2	<col:ac:unit>	real	10.2

Figure 9. Example Metadata Template File (*.mtf) contents.

[22] The contents of an *.mtf file corresponding to the *.mif file structure shown in Figure 6 are illustrated in Figure 9. It is similar to the *.mif but with information about how to format the *.mif file from a programming point-of-view. We use the *.mtf file to configure metadata editors and control the import/export processing of *.mif versions. It is the authoritative definition of a given *.mif version and can be processed to produce *.mif files automatically as well as to construct metadata catalogue relational databases tables. Note that the value ⟨ADO Name⟩ is used to indicate that that name of the appropriate ADO should be inserted for the value field in the *.mif file when it is produced. The *.mif contains information sufficient to unambiguously read and write the contents of the *.mif file and it explicitly defines the structure and contents of the corresponding type of ADO. This is a representative set of parameters and should not be construed as limiting but it is sufficient to construct the appropriate *.mif file for our example.

[23] Import and export filters can be readily built to read or write such ADOs and *.mif files and, in this way, digital objects and associated metadata get published in a standard and useful way while retaining the flexibility to modify the structure under community control over time as the science requires.

5.2. Naming Conventions for ADOs and *.mif Files

[24] So far in the example above we have used a simple file name as a prefix. In a practical system

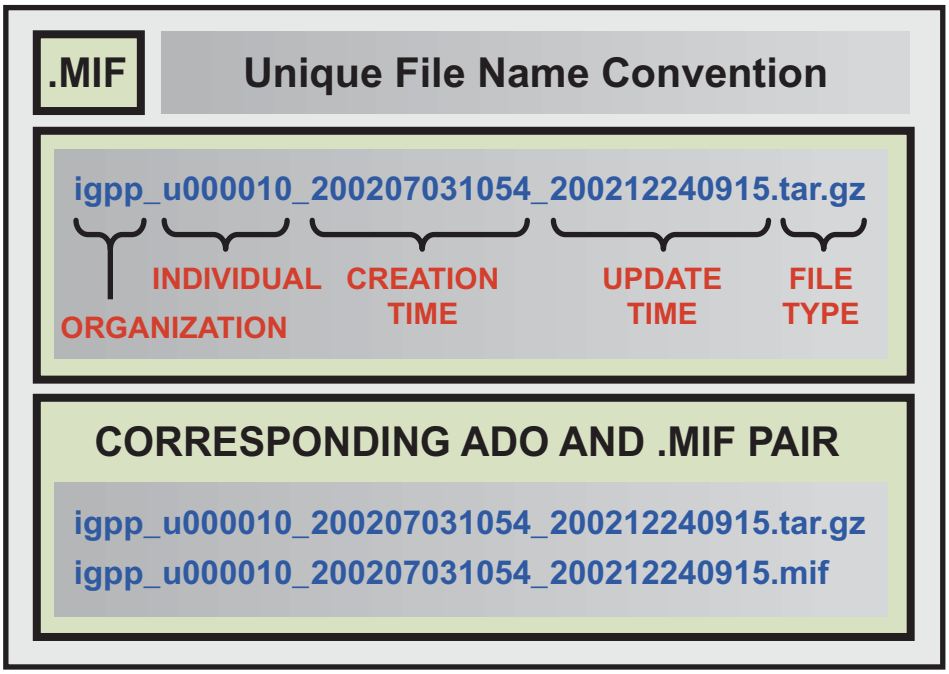


Figure 10. Naming convention for an ADO (above) and the correspondingly named ADO, mif pair. Here underscores separate parts of the ADO filename and are not related to the hierarchy referred to in Figure 7.

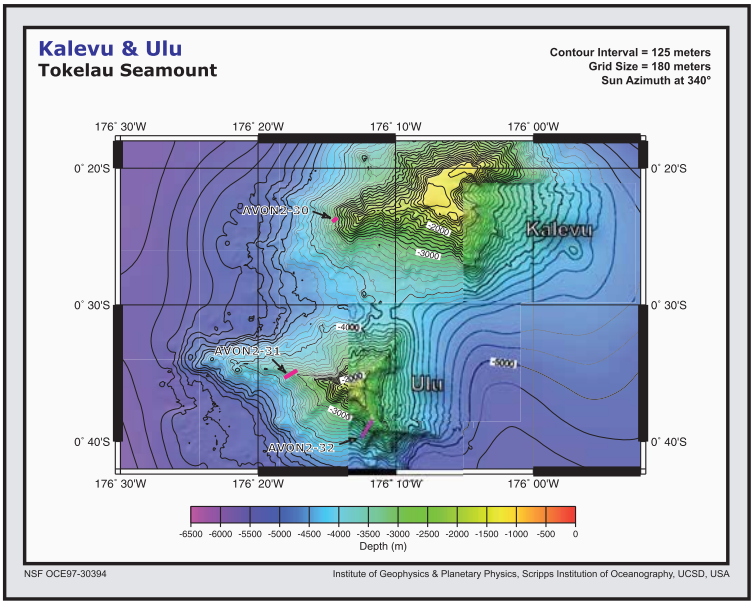
for interoperable data sharing, there must be a means of ensuring the unique naming of each pair of an ADO and its *.mif file. The reasons for unique naming include in particular the citeability and traceability of parameters. We must be able to uniquely identify data used in analyses and be able to refer to specific and unambiguous data objects in our work. One convention we have used successfully in a number of applications is shown in Figure 10.

[25] The creation datetime and the version datetime have the same value when the ADO and *.mif files are first created or published but subsequent versions of the data and metadata that may be issued for various reasons will update only the version date. In this way, the same basic ADO and *.mif files will sort together in a computer's file system making it easy to find the most current version of an ADO and *.mif file as well as the entire historical record of that ADO. We have put igpp in the front of the name to identify the organization that produced (i.e., published) the data. This particular combination of codings enables the ADO to be uniquely named within the collection of digital objects within the IGPP as well as without. Properly done, a scientific data network should have a community-based group that ensures that each data publisher is assigned a unique identification. This can be relatively easily done by using the above convention and by assigning unique codes for all participants in the data network.

[26] This approach is not limited to geochemistry data, nor to table data. For example, it is the metadata approach being used for the SIOExplorer project (http://SIOExplorer.ucsd.edu). An additional example of the application of the *.mif approach is presented in appendix A. This example illustrates the use of this approach to document raster images, especially maps, using the *.mif conventions.

6. Summary and Conclusions

[27] The ability of investigators to share data is essential to the progress of interdisciplinary and integrative scientific research. This is true even within individual disciplines. Here we have described a range of information architectures for effecting differing levels of standardization and centralization. We have proposed a generic .mif metadata interchange format for use in the earth sciences and potentially other disciplines and described how these can be created using a meta-



LABEL	METADATA PARAMETER	VALUE
MIF Version	MIF_Version	2.3.1
File Name	GEO_Geological-Maps-Sections_ADOname	kalevu-ulu.600.jpg
MTF Name	GEO_Geological-Maps-Sections_MTFname	agu.geo.map.mtf
Feature Name	GEO_Geological-Maps-Sections_FeatureName	Kalevu:Ulu
Region Name	GEO_Geological-Maps-Sections_RegionName	Tokelau Seamounts
Tectonic Plate	GEO_Geological-Maps-Sections_TectonicPlate	Pacific
Latitude S	GEO_Geological-Maps-Sections_LatS	-0.70
Latitude N	GEO_Geological-Maps-Sections_LatN	-0.30
Longitude W	GEO_Geological-Maps-Sections_LonW	183.50
Longitude E	GEO_Geological-Maps-Sections_LonS	184.13
Contouring	GEO_Geological-Maps-Sections_Contouring	val=125:unit=m
Grid Size	GEO_Geological-Maps-Sections_GridSize	val=180:unit=m
Sun Azimuth	GEO_Geological-Maps-Sections_SunAzimuth	val=340:unit=degrees
File Size	GEO_Geological-Maps-Sections_FileSize	val=79:unit=kb
Map Width	GEO_Geological-Maps-Sections_MapWidth	val=699:unit=pixels
Map Height	GEO_Geological-Maps-Sections_MapHeight	val=577:unit=pixels

Figure A1. This depicts a set of metadata parameter that are useful for map images that are also treated as ADOs. The image file is the ADO in this case.

data interchange template file that is controlled by the community and used to define the various types of ADOs and there corresponding *.mif files. While using a highly structured but openly available metadata format we are able to achieve effective data sharing across digital libraries, data archives, libraries and research projects. The most recent contributions to this discussion and .mif applications and examples may be found at http:earthref.org\metadata\GERM\.

Appendix A. Map Data Example

[28] This appendix depicts a set of metadata parameter that are useful for map images that are also treated as ADOs. Figure A1 is the ADO in this case.

References

Baru, C., R. Frost, R. Marciano, R. Moore, A. Rajasekar, and M. Wan, Metadata to support information-based computing

environments, paper presented at IEEE International Conference on MetaData 97, Nat. Oceanic and Atmos. Admin., Silver Spring, Md., 1997.

Daniel, R., and C. Lagoze, Extending the warwick framework: From metadata containers to active digital objects, *D-lib Mag.*, DOI 10.1045/dlib.magazine, 1998. (Available at http://www.dlib.org/dlib/november97/daniel/11daniel.htm)

Helly, J., New concepts of publication, *Nature*, *393*, 107, 1998.

Helly, J., T. T. Elvins, D. Sutton, and D. Martinez, A method for interoperable digital libraries and data repositories, *Future Generation Comp. Syst.*, *16*(1), 21–28, 1999.

Helly, J., T. T. Elvins, D. Sutton, D. Martinez, S. Miller, S. Pickett, and A. M. Ellison, Controlled publication of digital scientific data *Comm. Assoc. Comp. Mach.* (*CACM*), 2002.

Michener, W. K., et al., Nongeospatial metadata for the ecological sciences, *Ecol. Appl.*, *7*, 242–330, 1997.