Database Standards Case Study – Metadata Standards

Metadata is an important component of a database standard. For your review is a case study describing the development and deployment of a content metadata standard done by the USGS. Included in the paper are ideas and lessons learned during this process. While the data is related to marine scientific data, this example has limited direct applicability to the goals of our workshop. It is included primarily to spark the thought process related to the generation of a metadata standard and to make us aware of issues relating to the development of a metadata standard.

USGS case study highlights:

- Focus is on the ability to search and retrieve data
- Emphasis is on metadata standards but many concepts apply to real data
    - Examples: For geographical record location, is lat/long enough, or should state, region, etc. be accommodated?
- Establishment of specific vocabularies or value lists (flora/fauna, location names, for example)

Click for *Printable Version in PDF format*

Content Metadata Standards for Marine Science: A Case Study

USGS Open-File Report 2004-1002

by Rebecca L. Riall, Fausto Marincioni, and Frances L. Lightsom

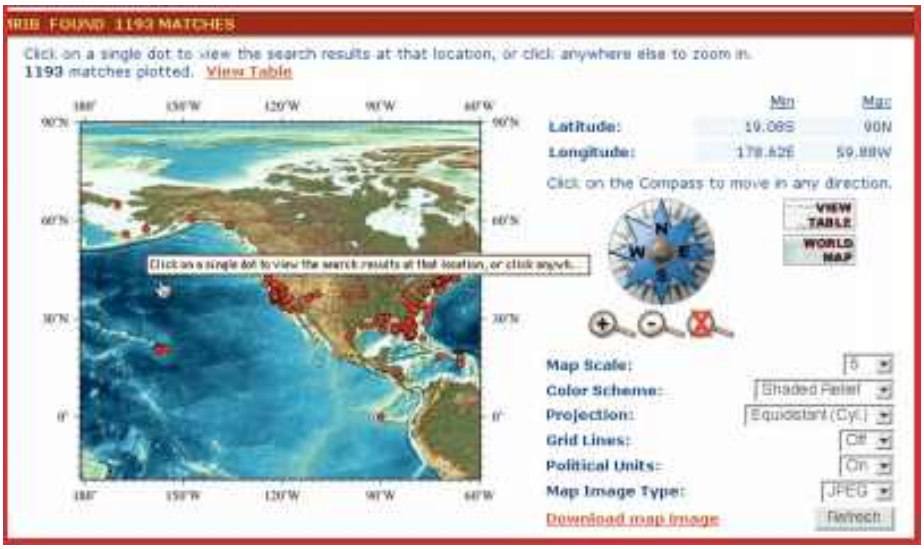


Figure 5. Map view and named locations are two ways the MRIB interface provides the spatial context of information resources. *Click on figure for larger image.*

Abstract

The U.S. Geological Survey developed a content metadata standard to meet the demands of organizing electronic resources in the marine sciences for a broad, heterogeneous audience. These metadata standards are used by the Marine Realms Information Bank project, a Web-based public distributed library of marine science from academic institutions and government agencies. The development and deployment of this metadata standard serve as a model, complete with lessons about mistakes, for the creation of similarly specialized metadata standards for digital libraries.

"Any use of trade names is for descriptive purposes only and does not imply endorsement by the U.S. Government."

Department of Interior / **U.S. Geological Survey** / **Coastal and Marine Geology**

Introduction

There is broad interest, for both educational and practical reasons, in understanding the processes of the coastal, marine, and lake environments through the scientific lens. (For brevity, the remainder of this paper will use marine to indicate all three environments.) Marine scientists are challenged to share their knowledge with coastal residents, government decision-makers, fishers, learners, and other lay people.

The rise of the Internet, particularly the World Wide Web, has made the sharing of scientific information—to both academic and lay audiences—an apparently instantaneous process. A scientist may author a Web page to share his or her data the same day that data is processed. (Or he or she may instruct a computer to generate the page automatically at regular intervals.) This paper will use the general term information resources to include Web pages (individual HTML documents which may have other media types embedded in them), Web sites (agglomerations of one or more Web pages), and other Web-served media which present scientific information at any level of technical difficulty.

Access to the wealth of scientific information available over the Internet is impeded because most Web searching tools (among them search engines such as the popular Google) do not provide an efficient way to find scientific information. One limitation of traditional search engines is that they cannot infer synonymous words. If one asks Google to find pages containing the text "Southeastern U.S.," it will not intuit that one also desires pages that substitute for "Southeastern U.S." the names of individual states constituting that region. Secondly, traditional search engines cannot rule out homographs and phrases that include the searcher's words, but are irrelevant for the searcher's purpose. For instance, a searcher wanting scientific information about Monterey Bay might run a search for the phrase "Monterey Bay." The results would likely omit relevant pages that refer to the area as "Monterey Sanctuary" (but not "Monterey Bay") while including irrelevant pages such as travel guides and menus for local restaurants. A third limitation of traditional search engines from the perspective of one seeking scientific information is that they do not provide any evaluation of scientific merit.

If search engines based on automated textual analysis fail to meet the needs of one who seeks specific scientific information, then what? Efforts like the Open Directory Project (http://www.dmoz.org) provide an alternative to traditional search engines (which often link to them), but also have their drawbacks. Such projects, which for convenience we will call "Web directories," however, also have their limits. For one thing, they usually will only place a single Web resource in a single classification. This one-page-one-listing strategy prevents flooding of the directory by particular Web sites of especially broad scope. Thus, a Web site whose pages represent the words taken from a dictionary is not guaranteed an entry in every category of the directory. On the other hand, the one-page-one-listing strategy also ensures that Web sites will only be listed in the most general terms, and prevents searching for information resources by more than one criterion. A second problem is that the form of Web directory listings is simplistic, providing

the searcher with only a title, a URL, and a brief description, so the user has little information at hand to compare listings, and thus must "surf" to find the most useful resources, if any.

Specialized Web directories—for instance, annotated lists of links gathered by a specialist in a particular field—may offer more up-front information, but they too have their drawbacks. They cannot usually aggregate information between scientific fields. If they are maintained manually, then they are limited both in the number of information resources they can describe and in the precision by which these resources are differentiated. They cannot interchange data with similarly-functioning directories; they do not use metadata.

The U.S. Geological Survey needed a better way to present its coastal and marine geology Web resources as an organized collection, and to eventually integrate resources from other agencies, so it developed the Marine Realms Information Bank (MRIB) project. The MRIB, which can be found at http://mrib.usgs.gov, is a Web-based distributed library about marine environments. By calling it a library, we mean that it catalogues information resources in a consistent, rigorous way, just as a library catalogue does, and permits searching of the catalogue. Moreover, it attempts to duplicate the most basic forms of "good advice" that could be offered to a searcher by a knowledgeable reference librarian, using detailed descriptions of information resources to suggest potential avenues for further searching. By qualifying the "library" as "distributed," we mean that the information resources are not held in the MRIB. They are described in the MRIB, in documents that remain in the MRIB catalogue, but the resources themselves are elsewhere. A central part of the MRIB is an ontology for Web materials about the marine environments. That ontology is what sets the MRIB apart from search engines and Web directories. The ontology, which is an abstract organizational schema, is manifested by a metadata standard which is the focus of this paper.

As suggested earlier, information resources listed in the MRIB catalogue remain on their originating servers (and in the control of their creators), while the MRIB stores and searches locally-held metadata that describe those resources. The MRIB currently catalogues information resources according to an ontology constituted by thirteen facets (metadata fields with controlled vocabulary lists) as well as a suite of other textual and numerical metadata fields that do not rely on controlled vocabularies. The MRIB's metadata fields and vocabularies have been tailored to adequately describe the distributed library content (coastal, marine, and lake science documents) while serving a broad audience spectrum ranging from elementary school students to policy makers to oceanographers.

In this paper, we will outline the special challenges of categorizing digital information about the marine realms for a heterogeneous audience. Then we will review several metadata standards and their usefulness in light of those challenges. Next we will describe the process and outcome of the MRIB metadata standard development. Finally, we will evaluate the suitability of both the metadata standard and its development process to meet those challenges.

Back to Top

Title / Introduction / Cataloguing / Evolution / Case Study / Discussion / Conclusion / References /

Department of Interior / **U.S. Geological Survey** / **Coastal and Marine Geology**

USGS Privacy Statement / Disclaimer / Accessibility

This is http://woodshole.er.usgs.gov/drafts/accessib/meta_standards/html/intro.html
Maintained by webmaster-woodshole@usgs.gov
Modified Tuesday, 02-Mar-2004 11:11:21 Eastern Standard Time

Challenges in Cataloguing the Marine Realms

Organizing and aggregating information about marine environments is a necessity, but it is not trivial. Six specific challenges were encountered during the MRIB's development. It is important to note that several of these challenges arose from conscious decisions made by the MRIB developers, and would not necessarily be faced by any digital library project. Information about the difficulty of responding to each challenge may prove useful in making selections about the scope and audience of other digital library projects.

Perhaps the most difficult challenge is posed by the MRIB's broad potential audience: to meet the needs of many groups and individuals, the MRIB developers strove to organize information in ways that would anticipate and meet many users' search strategies. Because no search strategy could ever be sufficiently universal to assist to every possible user, the broad-audience challenged mandated many avenues to the same information. It should be noted that the MRIB does not attempt to meet every possible strategy, but only to provide a large set of possible strategies. This audience consideration strongly tempered consideration and approaches to the other challenges.

The second challenge is that much research on the Earth is firmly located in specific places and times, and research on Earth's water bodies is no exception. Even so simple a measurement as water depth may vary greatly within a few lateral feet and a few hours, so spatial and temporal placement of data is crucial to underwater studies. Thus it is important that any catalogue dealing with the marine environments allow classification and searching of information resources by location and time.

A third challenge is that "marine science" encompasses an exceptionally vast range of academic disciplines: all the natural, physical, and social sciences are concerned with standing-water environments. The humanities, the practical fields (such as education and law), and the arts also have something to say about the marine environments. Thus, a truly inclusive ontology (categorization system) for marine science must appeal to the disparate mindsets and internal organizations of many disciplines. Developers and users of this ontology must remain aware that terms from one discipline may represent wholly different concepts in a another discipline. The MRIB has simplified the problem of excessive disciplinary scope by limiting its content to the scientific and educational disciplines for the near future.

Because the MRIB's ambition is to serve users who are not necessarily scientifically trained, the ontology and term lists must not be over-full of jargon (this is the fourth challenge). Avenues must exist for the scientist who maps the seafloor, the student who is reporting on cephalopods, and the concerned citizen whose home is in a flood zone to each find the information she or he requires. It is an ongoing process for the MRIB ontology to balance the conflicting needs for 1) precision of terminology; 2) consideration of conflicting term connotations between academics and lay people, as well as between academic disciplines; and 3) avoiding excessive jargon.

Although not specifically associated with cataloguing the ocean and lake

environments, a further challenge posed to the development of a marine science ontology is the need to moderate the number of search fields it offers. One limitation of traditional electronic library catalogues is that they rely on simple lists of authors, subjects, and titles, which provide little or no contextual information about related subjects and items. The MRIB provides this information in a more organized set of hierarchical topics, so it requires more than those three basic metadata fields. However, too many fields or fields that overlapped too greatly could overwhelm the user and render an intuitive interface impossible.

Those who are perhaps best able to precisely and elaborately catalogue documents are the authors and maintainers of those documents. Often these people are already accustomed to negotiating the murky waters of interdisciplinary and educational communication. Although an MRIB record about a document may be created by someone uninvolved in the document's production, the MRIB categorization scheme is intended to encourage authors to create their own bibliographic records. Self-cataloguing also has the benefit of involving authors in the developing of the scheme, where their suggested terms may be vital. Designing an ontology which encourages author generation of metadata records was a final challenge to MRIB development.

In summary, the most significant challenges posed in creating a categorization scheme for a widely-usable digital library of the ocean, lake, and marine environments are:

- Accommodate geospatial and temporal "footprint" of information;
- Integrate information from a broad spectrum of academic disciplines, while minimizing discord from different term usage across disciplines;
- Organize information so that a variety of searching strategies can succeed;
- Minimize jargon;
- Use enough metadata to provide the above, while remaining cautious about the total number of searchable metadata; and
- Encourage composition of metadata records by document authors.

Back to Top

Title / Introduction / Cataloguing / Evolution / Case Study / Discussion / Conclusion / References /

Department of Interior / **U.S. Geological Survey** / **Coastal and Marine Geology**

USGS Privacy Statement / Disclaimer / Accessibility

This is http://woodshole.er.usgs.gov/drafts/accessib/meta_standards/html/challenge.html
Maintained by webmaster-woodshole@usgs.gov
Modified Tuesday, 02-Mar-2004 12:09:51 Eastern Standard Time

The Evolution of Metadata Standards Relevant to the Marine Sciences

A Brief History of Metadata

Traditionally, library cataloguers and archivists have used the term metadata to refer to descriptive information used to index, arrange, file, and improve access to a library or museum's resources (Gilliland-Swetland, 1998). This use of metadata follows from the Greek etymology of the prefix meta, which literally translates as "with, among, after, or behind." Thus "metadata" suggests something "accompanying" data but not essential to it. Recent advances in information technology and the rapid emergence of the digital library have somewhat altered the perception of the term metadata among information managers; metadata are no longer auxiliary definitions or descriptions of some library resource, but a fundamental dimension of said resource.

Because the digital library field is young, its terminologies and concepts are often defined vaguely or contradictorily between authors. The word metadata has met such a fate, and many definitions of it have been invented, refined, and circulated. Numerous examples could be cited here, but what is important is that — whether in its traditional context or in a digital library context — the key purpose of metadata remains the same: to facilitate and improve the retrieval of information.

An early use of metadata in the digital world occurred in the 1960's, with the advent of the international Machine-Readable Cataloguing (MARC) standards and the Library of Congress Subject Headings (LCSH). These standards were used to develop automated retrieval systems such as Online Public Access Catalogs (OPAC).

Back to Top

Metadata Standards Relevant to the Marine Sciences

A digital library for marine science, indeed for any Earth science, has one primary need beyond those of more general libraries: to describe the spatial and temporal coordinates associated with information. Although the MRIB is intended to enable both browsing and searching, it would certainly be possible to build a marine science digital library that relied on more standard bibliographic data and a more traditional library catalogue, oriented toward searching rather than browsing of categories. Such a library would trade browseability for truncated development time. Following is a discussion of some common metadata standards which are especially relevant to cataloguing marine resources, their advantages, and their shortfalls in relation to the MRIB goals specified above.

Back to Top

o MARC21, Machine-Readable Cataloguing

The Library of Congress developed the Machine-Readable Cataloguing (MARC) format in the 1960's to aid librarians in computerizing their catalogues and sharing records with one another (Furrie, 2000). MARC, presently in its twenty-first iteration, uses character codes to name bibliographic data fields (such as "100 1# $a" for its "Author" field). It was responsible for the computerization of library catalogues over the past several decades. Although efforts have been made to adapt MARC to electronic materials (Library of Congress, 2002), it still has notable disadvantages, including its human-unfriendly field names, inability to describe computer formats precisely, and age. These limits combined with MARC's inability to handle numerical spatial data make it inappropriate for an MRIB-style digital library, especially one which encourages authors to create their own bibliographic records. Some members of the traditional library world have also begun to reject MARC in favor of XML bibliographic records, which are expected to ease the integration of paper and electronic resources (Miller, 2000).

Back to Top

o Federal Geographic Data Committee Content Standard for Geospatial Metadata (FGDC-CSGM)

The FGDC began drafting its Content Standard for Geospatial Metadata in 1992 (Federal Geographic Data Committee, 2000). According to the FGDC-CSGM Workbook, this standard is intended to facilitate three uses of data:

> "(1) to maintain an organization's internal investment in geospatial data, (2) to provide information to data clearinghouses and catalogs, and (3) to provide information needed to process and interpret data transferred from another organization."

FGDC-CSGM is used by many clearinghouses of data because of its thoroughness and its ability to describe data in very precise terms. However, the FGDC-CSGM is so specific that it becomes unwieldy to apply, and thus is undesirable for a catalogue of Web-based materials which are not necessarily raw data and for which much of the information required by the FGDC-CSGM may not be readily available or desirable for searching and browsing. Moreover, to use FGDC-CSGM metadata, one practically needs to be a specialist in the standard. This is not ideal for a library of Web content such as the MRIB that encourages Web document authors to compose the metadata profiles for their own documents. Nor do many of the FGDC fields possess controlled vocabularies, whose absence which makes FGDC records less interoperable for searching. For instance, there is no FGDC authority list of author names, so there is no certainty that a search of a collection of FGDC records for an author name will find all the records linked to a particular author (who might sometimes use initials rather than a full given name, or who might change his or her surname). For these reasons the MRIB team chose to develop a simpler, more focused (but still very detailed) metadata standard that would facilitate record interoperability, rather than designed with the meet-all-possible-needs approach of the FGDC standard.

Back to Top

o Dublin Core Metadata Initiative (DCMI)

The DCMI emerged from a 1995 workshop during which participants discussed

essential categories by which Web resources could be catalogued (http://dublincore.org/about/history/). The present-day DCMI provides a set of standard field-names with the aim of "making it easier to find information," the slogan on the project's Web site (http://www.dublincore.org). DCMI specifies syntactical structure for various elements (such as the contents of the field, the controlled vocabulary from which the contents were derived, etc) of each field. With one exception, DCMI does not provide controlled vocabularies for metadata fields; instead, it registers such controlled vocabularies and allows metadata cataloguers to use (and to specify in their metadata) a relevant vocabulary developed elsewhere (Dublin Core Metadata Initiative, 2003a; DCMI). The exception is a rudimentary, flat controlled vocabulary for the DCMI "Resource Types" field (including such terms as "Image," "Event," and "Sound") (DCMI, 2003b). The field names in Dublin Core are human-language terms like "Publisher," rather than MARC-style machine codes. Because DCMI was developed specifically for electronic resources, it dispenses with some of the extraneous bibliographic fields that are irrelevant to electronic resources. Moreover, because it does not focus on describing "data" in the rawest sense, DCMI is simpler than FGDC and more broadly applicable. The DCMI provides fields to specify time ("Period") and location ("Points"), both of which are crucial to describe information resources about the Earth.

Back to Top

o ADEPT Metadata Standard

The ADEPT metadata standard results from collaboration among NASA's JOINed Digital Library, the Digital Library for Earth System Education (DLESE), and the Alexandria Digital Earth Prototype (ADEPT) . ADEPT, which is still in development, promises great versatility in dealing with Earth sciences resources in general. In particular, because the projects involved include a library of mostly raw data (Alexandria) and a project to organize kindergarten – , through college-level educational resources (DLESE), ADEPT will need to find effective ways to sort information by technical level. The ADEPT standards, being specialized for the Earth sciences, include fields competent for describing space and time in several ways. Although not strictly part of the ADEPT metadata, the Alexandria project has also developed an extensive polygon-based gazetteer which, in conjunction with geospatial metadata specified by the ADEPT standard, may provide very accurate location searches. That said, because ADEPT is adapted for the broader Earth sciences, it has some limitations in the scope of the marine sciences. From those sections of the ADEPT standard that are publicly available, it is difficult to judge how well ADEPT will describe information outside the disciplines of geology, geography, and education. The current standards propose a section tailored to the metadata needs of specific disciplines, but no details about the fields in that section or the breadth of disciplines covered are yet public.

Back to Top

Title / Introduction / Cataloguing / Evolution / Case Study / Discussion / Conclusion / References /

Department of Interior / **U.S. Geological Survey** / **Coastal and Marine Geology**

USGS Privacy Statement / Disclaimer / Accessibility

This is http://woodshole.er.usgs.gov/drafts/accessib/meta_standards/html/evol.html
Maintained by webmaster-woodshole@usgs.gov
Modified Friday, 06-Feb-2004 09:00:04 Eastern Standard Time

USGS

Content Metadata Standards for Marine Science: A Case Study, USGS Open-File Report 2004-1002

The MRIB Metadata: A Case Study

NOTE: *Library card catalogues (drawers of cardstock slips, each bearing subject headings and other vital data about a particular volume) were once ubiquitous symbols of information organization in the physical world. Because the MRIB is drawing on the concepts of these paper-and-ink metadata records, its metadata records are called "Electronic Index Cards," or EICs.*

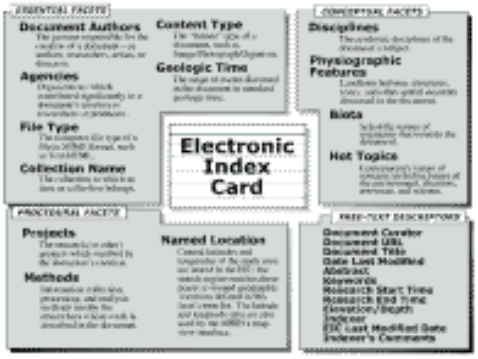
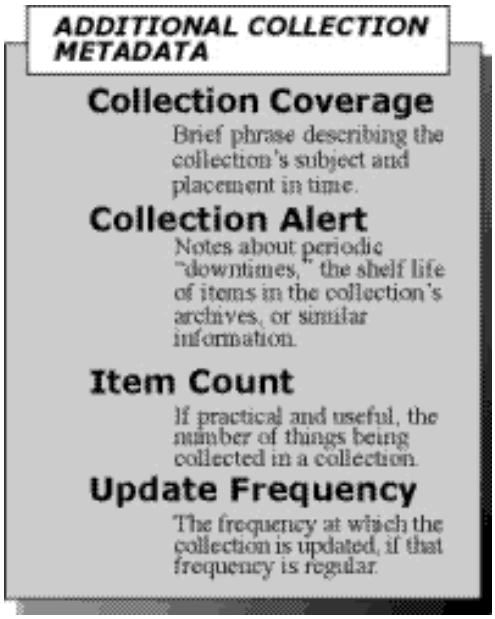


Figure 1. The MRIB's item metadata fields. *Click on figure for larger image.*

At the heart of the MRIB project is the creation of a metadata standard specialized for representing the marine sciences and for providing access to users of varied technical competence. The discussion that follows will address how this standard has developed and how some of its early mistakes are being corrected. The MRIB case study may serve as an example of how the challenges of specialized-content digital libraries may be met, and the account of pitfalls along the way may be useful in the establishment of other such libraries. More technical information about the actual implementation of the MRIB metadata, such as the computer format in which EICs are stored and how the interface functions, can be found in Marincioni and others (2003). The complete controlled vocabulary lists, metadata dictionaries, a DTD (Document Type Definition used to guide and validate eXtensible Markup Language, or XML, documents) for coding MRIB records in XML, and other supporting documents are stored on the MRIB server at http://www.mrib.usgs.gov/controlled_vocabulary/.

Although the non-MRIB metadata standards outlined earlier have evolved substantially since the inception of the MRIB, the MRIB has unique metadata needs by virtue of its subject matter and the kinds of resources it catalogues. Metadata fields for the MRIB were required to address the particular needs of Web-based information resources about marine science. The fact that marine science is a broad spectrum of endeavors —which includes work by educators, anthropologists, and historians, in addition to the more obvious natural scientists (such as oceanographers and geologists) — complicated the creation of these metadata fields. Thus the MRIB required a new metadata standard



(described in Figure 1 and Figure 2), rather than re-use of other standards. The keystone of this new standard is use of controlled vocabularies wherever conceivably helpful. A controlled vocabulary, that is, the complete list of valid terms for a given facet, facilitates both finding and cataloguing because it ensures that a concept will always be assigned the same classificatory term within an encompassing facet. Additionally, since the set of possible terms within a facet is known, rather than undefined, relationships between terms can be explicitly defined. Hierarchical term relationships are emphasized in the MRIB because they allow the user to adjust the level of specificity for her searching or browsing. Example terms from the controlled vocabulary of each of the item metadata fields appear in Table 1.

Figure 2. The additional metadata fields for collections. *Click on figure for larger image*.

Table 1: Examples of Terms from the Controlled Vocabularies of the Facets

ESSENTIAL FACTS

FACET NAME	EXAMPLE
Document Authors	Riall, Rebecca L. rriall@usgs.gov
Agencies	Academic Institutions/United States of America/Indiana/ Indiana University/Bloomington
Content Type	Images/Still/Photographs/Ships and Other Platforms
Geologic Time	Phanerozoic/Cenozoic/Tertiary/Neogene/Miocene
Collection Name	Mobile Bay Satellite Images
Collection Title (collection metadata only)	Gulf Coast Satellite Images/Mobile Bay Satellite Images

PROCEDURAL FACETS

FACET NAME	EXAMPLE
Projects	Marine Realms Information Bank
Methods	Field Observation/Remote Sensing/Aerial Photography

Location	(Numerical latitude and longitude are recorded for the study area of the document. The search engine matches these points to named locations that are defined by bounding ranges. Example below is of a named location.)Seas and Gulfs/ Americas/North America/Gulf of Maine/

CONCEPTUAL FACETS

FACET NAME	EXAMPLES
Disciplines	Geology/Sedimentology
Physiographic Features	Landform/Islands/Barrier Islands
Biota	Eukaryota/Metazoa/Annelida/Polychaeta
Hot Topics	Environment/Environmental Issues Involving Sediment/ Sediment Interaction with Pollutants

Back to Top

Items and Collections

In a physical library, it is customary to catalogue items based on their material presence —a self-contained three-page pamphlet would have a record in the catalogue, while an article of similar length, but contained by a journal, would not. Such "sub-items" as individual articles in a single bound journal would instead be included in subsidiary indices —in this case the periodical index provided by the journal publisher, a separate searchable catalogue of the journal, or a single-discipline catalogue (such as Georef). However, in a digital library, the choices about cataloguing depth, or granularity, are less fixed.

The complexities of granularity choices are perhaps best illustrated by example. For instance, a Web site might host a collection of one hundred seafloor images, with two images on each page. This leaves several possibilities for what components of the whole Web site will be catalogued: 1) Only the main page, which indexes the pages containing the actual images, is catalogued (one record); 2) Each HTML page, with its two images, receives an individual catalogue record (a total of fifty records plus the index record); or 3) Each image receives an individual catalogue record (a total of one hundred records plus the index record). Each of these three selections has its advantages and disadvantages. If only the main page is indexed, then a user who is interested in a very specific location, such as offshore Alabama, would be unable to find the image of that area directly through the catalogue records. On the other hand, if individual HTML pages are catalogued, a searcher who wants to find larger seafloor images, such as an image of the entire Pacific Ocean, might be inundated with irrelevant findings. Cataloguing individual images rather than individual HTML documents could further exacerbate the deluge even as it permitted direct access to the images from the catalogue. One might term the extreme of only one record "very low granularity" and the extreme of 101 records "very high granularity."

The first versions of the MRIB followed a moderate approach toward granularity. In this approach, the MRIB team applied a simple rule: it was acceptable to catalogue any Web page or PDF file that had at least two self-contained items of information. For instance, a photo and a descriptive caption would qualify, while the photo alone would not. The two-items rule served as a metric to eliminate items that had little or no value independent of their context. For example, it precluded the cataloguing of a page that listed only the definition of a single term or an image that was presented with no information about its importance.

The success of this tactic depends partly on the initially-small collection size of some 3,000 records. However, as the MRIB collection has grown, the flaws of this "moderate" approach have become apparent. The chief problem is that is possible for some user searches to return a large number of highly-similar pages. For example, a set of records for 2,000 DODS (Distributed Oceanographic Data System) resources, each with nearly the same metadata profile, flooded the MRIB with difficult-to-distinguish results when they were temporarily introduced. On the other hand, eliminating individual item records from the MRIB in favor of a one-record-per-collection approach would also make valuable items difficult to locate directly from the catalogue.

Two separate sets of metadata records constitute a newer version of the MRIB metadata catalogue —one set for collections of resources and another for individual resource items. An "item" is an individual document (which may combine several different information types, such as images and text, that are viewed as a single unit) such as an HTML page, PDF file, or plain-text file. A "collection" is typically a set of pages intended by their author or authors to form a cohesive aggregate, such as a Web site (collection of Web pages). However, a "collection" might also be a single document on which some fundamentally similar things are collected (such as a single HTML page which includes a series of hundreds of photographs of marsh organisms).

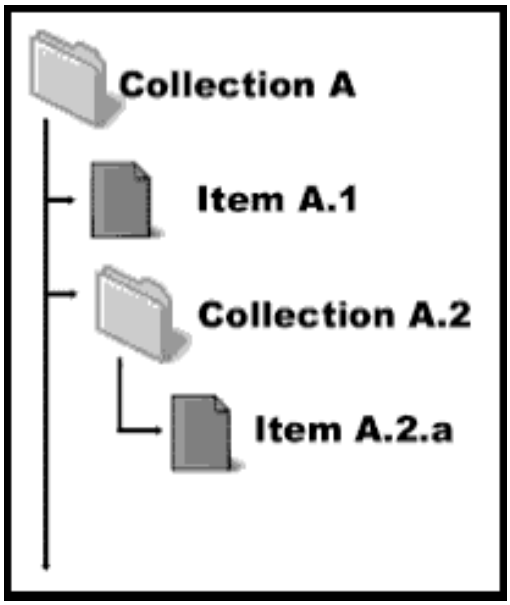


Figure 3. Item and collection "nesting." *Click on figure for larger image.*

As part of the newer dual-granularity catalogue, separate item and collection metadata profiles have been specified to describe characteristics unique to each record type and to define hierarchical relationships between 1) items and collections, and 2) collections and their encompassing collections (collections of collections), as illustrated in Figure 3. The MRIB user will be able to choose whether to search only collections or to search individual items. Both metadata profiles include the same facets and fields for subjects and standard bibliographic descriptors. The details of the item and collection profile types and the rationale by which they developed are described below.

Back to Top

The MRIB Item Metadata Fields

The MRIB is intended to be a "browsing engine" as well as a search engine.

Figure 4. Controlled vocabulary for one facet, Physiographic Features, in action. *Click on figure for larger image*.

In fact, the earliest version of the MRIB did not permit searching for user-defined strings, but rather relied wholly on point-and-click browsing of the term lists. The intention to support browsing dictates that possible search subjects in the MRIB must be organized in some logically consistent way. The MRIB team chose to organize concepts into groups of related subjects, with each group containing broader, narrower, and co-level terms. Each top-level group representing major classification criteria (a facet) was assigned its own controlled vocabulary (see Figure 4).

The valid terms for a given facet, constituting its controlled vocabulary, are listed in a term list specific to that facet. Each facet's term list is structured in a database-readable format. The lists store additional information about each term, including a brief definition of the term as it is used in the MRIB and book-keeping data (such as who entered the term, when it was entered, when it was last modified, and whether it has been approved by the MRIB team). The term definitions are especially important, because the terms do not simply record words used in the cataloged documents, but instead represent concepts that, to be precisely defined, often require sentence-length explanations. Explicit term definitions encourage the consistent application of terms that otherwise would be ambiguous because of cross-disciplinary, regional, situational, or other differences in use. It is unavoidable that disciplines will use terms in radically as well as subtly different ways. Similarly, it is unavoidable that scientific and nonscientific users will also use words differently. Definitions of terms as they are used in the MRIB are thus necessary, and may encourage users to reflect on different uses of a word.

The current MRIB item metadata fields can be sorted into four major groupings (Table 1). The first group is the essential facets (Document Author(s), Agencies, Collection Name, Content Type, File Type, and Geologic Time) which are considered the minimal required facets for cataloguing any document (in addition to some fields lacking controlled vocabularies which are described below), because any resource will have these attributes. Authors and Collection Name may, on very rare occasions, be exempted from requirement. A second group contains the procedural facets — Projects, Methods, and Location (although it has a controlled list of terms, location terms that apply to a document are derived from numerical latitude and longitude values rather than by named locations embedded in the document's metadata record) —that describe the research processes which resulted in the document's birth. The third group encompasses the conceptual facets (Disciplines, Physiographic Features, Biota, and Hot Topics) that note the subjects of the document.

The fourth group consists of the free-text descriptors. These are metadata fields that, unlike the facets, lack controlled vocabularies. They are URL, Title, Document Curator(s), Document Last-Updated Date, EIC Indexer, Date of EIC Creation, Date of Last EIC Update, EIC Creator's Comments, Description,

Research Start Date, Research End Date, Other Keywords, Elevation (mean, maximum, and minimum values), Latitude and Longitude (mean, maximum, and minimum values; these metadata are also used at the interface level by the Location facet). These metadata provide additional (and in some cases, such as Title, critical) information to the MRIB user, but will naturally vary so greatly that controlled vocabularies for them would not be sensible. Of the free-text metadata fields, URL, Title, Document Curator(s), EIC Indexer, Date of EIC Creation, Date of Last EIC Update, Description, and Other Keywords are always required. It should be noted that Other Keywords stores any terms relevant to the document that cannot be described elsewhere in the EIC.

For instance, the common name of an organism, a geographic name which may not yet be defined in the Location facet, or other terms.

Back to Top

o Development of the MRIB Metadata

The present MRIB structure has evolved from its initial conception based both on the indexing-to-uncover process (which will be explained shortly) and on feedback from users. The original facets in the MRIB metadata standard were Author, Project, Discipline, Issue, Method, Geologic Time, Location, Class, Format, and Agencies. With the exception of Author and Project, the term lists for each facet were seeded with terms commonly used in USGS publications, the NASA Global Change Master Directory, and the Marine Biological Laboratory's Web site. New terms are added to Author and Project ad hoc. Although it might seem simpler to leave these two facets uncontrolled, the use of controlled vocabularies for them makes it more likely that, for instance, an author's works will appear under only one name, regardless of whether some publications give only the author's initials while others use the full name, or whether an author undertakes a legal name change. Similarly, projects may be variably referred to by official titles, acronyms, and informal names, so it is important that only one be used by indexers for a single project.

After this initial seeding, the MRIB EIC collection and controlled vocabulary lists were expanded using an indexing to uncover process, as described by Marincioni and others (2003). This process involves the experimental cataloguing of appropriate marine science information resources (thus creating new EICs) while maintaining lists of new concepts and the terms used by resource authors to describe them. These lists were eventually incorporated into the formal term lists. Although emphasis was placed on adding recurring concepts to the term lists, concepts that had not yet been seen to recur during the indexing were sometimes added because of their likelihood to recur in the future, their interest level, or other factors.

The terms themselves were derived, where possible, from the information resources themselves; when several terms existed for a concept (or when a concept crossed between facets) the least ambiguous and most widely-used terms were privileged. Redundancy was permitted between facets and between trees within each term list, so that a concept might appear in different aspects, in different facets. For instance, the basic concept of sediments appears in the Discipline facet as "Sedimentology" and in the Physiographical Features facet as "Soil." Additionally, the small sets of original terms included in the seed lists were often modified to better reflect common usage or to be more precise.

At times, the basic structure of the categorization scheme itself was significantly

revised. In some instances, new facets were created (Biota, Physiographic Features, and Audience); in others, the hierarchy of a facet was totally revised to permit a more consistent and scalable organization. Care was taken to favor revision, rather than to allow the ontology to become mired in trial forms that proved unsatisfactory for the practical tasks of organizing and locating information resources.

The organizational structures that emerged from the indexing-to-uncover process are flexible. New terms can be added easily, and after much indexing the broad categories within most of the facets have become optimized for expansion. It is expected that new methods and new concepts will continually emerge in the marine sciences and related disciplines, so the flexibility is a necessary feature. Certainly at present, new terms are frequently added to the lists.

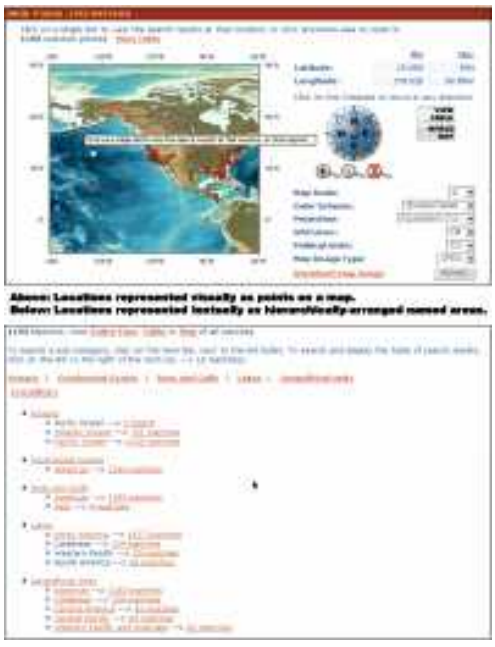
Back to Top

o Time and Location

It is reasonable to ask how the MRIB team has attempted to especially suit the MRIB metadata standard to marine science. One way is that, through the indexing process discussed above, terms and concepts have been gathered from marine science documents generated by experts, but their inclusion was also tempered with feedback from non-experts. Secondly, Earth science information is usually associated with two major concepts which the MRIB metadata have been made to emphasize: time and location.

Time in the Earth sciences is commonly expressed as a series of large-scale geologic blocks, ranging from the present backward to the Pre-Cambrian. The MRIB Web interface sorts items based on their ranges in this geologic time scale, which are entered in the EICs in the Geological Time facet. Because marine science also deals with significant changes, such as coastal erosion, over smaller-than-geological periods, such as years and decades, the MRIB standard includes metadata specifying data collection dates. However, these metadata do not have controlled vocabularies and will be discussed further later.

The MRIB interface provides two means to find resources by geographical location: Named Locations (a list-based interface) and Map View (which plots rough locations of resources on an interactive map), shown in Figure 5. Both of these means exploit the numerical latitude and longitudinal values stored by the MRIB metadata fields. Named Locations is considered a facet because it has a fixed (although expandable) list of locations, but, unlike the other facets, its values are not stored directly in metadata records. Instead they are derived from the geographical coordinates which are stored in the records. For instance, an EIC for a resource about Lake Michigan would record the four geographical coordinates of a box around the lake. The MRIB interface would both plot the



center of that box in Map View and list the resource as a match for all the Named Locations which overlapped the Lake Michigan bounding box (including Lake Michigan, North America, Indiana, and so on).

Figure 5. Map view and named locations are two ways the MRIB interface provides the spatial context of information resources. *Click on figure for larger image*.

That said, there are some problems with matching records to the Named Locations term list. Although its definition of locations by four bounding lines (on the latitude and longitude grid) is simple, the resulting bounding rectangle may become misleading, particularly for large, irregularly shaped geographical features. For instance, a bounding rectangle that encompasses all of the Pacific Ocean also includes much of the Atlantic. This means that, among other misleading results, a record with its central point in Cape Cod will be listed as a "match" for the Pacific Ocean.

Fortunately, one benefit of storing the coordinate data rather than named locations in the records is that these coordinates may be used by newer interface methods and different location lists without requiring modification of the metadata in the EIC. For instance, the Pacific-Cape Cod problem will be remedied in a future MRIB interface revision which permits including bounding polygons in the term list (the solution will not require changes to the metadata database). Thus the problems are with the interfaces to the metadata, and the metadata fields themselves will be compatible with more sophisticated interfaces.

Back to Top

o Major Revisions Along the Way: The Audience, Class, and Format Facets

As mentioned earlier, the categorization scheme was revised when needed during the early development of the MRIB metadata. Eventually, cataloguing and user interaction brought to light essential problems with the Audience, Class, and Format facets. These problems may be useful lessons for the development of other specialized categorization schemes.


Figure 6. Term lists from the old Facet, Class, and Audience facets.

AUDIENCE	FORMAT	CLASS
Disaster_Management	Atlas	Data
Model_Implementation	Data_Set	Derived_Products
Policy_Making	Data_Set:DODS	Knowledge
Public_Awareness	Database	Productions
Recreation	Dynamic_Id	Scientific_Publications
Resource_Management	Email_List	
Teaching	Equation	
Equipment_Use	FTP	
	GIS	
	Graph	
	Image	
	Map	
	Mosaic	
	Movie	
	Newsgroup	
	Photograph	
	Poster	
	Software	
	Software Applet	

These Audience, Class, and Format facets began with short, fixed term lists (Figure 6). All three suffered from both opaque individual definitions and overlap of purpose. Each of them attempted, from a slightly different angle than the others, to provide information about how a document might be used. However, it is simply not possible to objectively determine each way a document might be used. This is perhaps best clarified by example; we will consider the Audience term "Educator." For some documents, such as one titled "Fourth Grade Lesson Plan on Wetlands," the audience for that document might objectively be noted as "educators." But what of something like an instruction list for using a sonar mosaic processing program? This, too, might serve a

Figure 6. Term lists from the old Facet, Class, and Audience facets. *Click on figure for larger image*.

specialized "educator" well. Nearly any document might be reasonably expected to serve some kind of educational needs. In an attempt to address this kind of problem-which might be called the problem of explicit versus implicit purposes-the Audience facet was intended only to be used when there was an explicitly-stated audience for a resource (for instance, some resources, such as lesson plans, clearly were intended for teachers, so these would get the "Educator" term). Despite this precaution, early user test groups complained that they thought Audience was an effort to pigeonhole users and limit their browsing choices.

Class proved even more problematic than Audience. Based on interaction with scientists at the USGS, the MRIB developers were aware of scientists' wishes to be able to handily limit their MRIB searches to EICs that described raw data. An early MRIB model of "data" and "not data" posited that documents might be objectively classified by their level of removal from that "pure data" state: they might be data, they might be products derived from data, they might be knowledge synthesized from data analysis, or they might be predictions based on the three preceding stages. Although this model seemed heuristically useful, in practice the heterogeneous, contextual nature of Web documents and the variation of the human mind rendered it useless. For instance, a geological map alone might be considered the derived product of raw data-yet it also, to the extent that it extrapolates beyond the finite set of data points described during mapping, is a visual representation of predictions. Then again, that map might be considered base data for a synthetic map which combines small geologic maps into a larger-area map. Even if the map's creators intended the latter use, would it still be appropriate to rule out the map's other potential uses (as a derived product or as a prediction)? Early on, the MRIB developers realized they could not agree on how to classify some documents. Despite this, the field continued to be applied-but it was no surprise when users, too, found they could not understand how or what the Class of a document meant.

The MRIB team's experience with the Class facet suggests that developers of categorization schemes heed cataloguers' experiences; if cataloguers cannot agree on how to consistently apply a category, that category is likely to pose problems for end-users as well. This is not to imply that a category and its applications will ever be universally agreed-upon, only that it should be agreed upon with some consistency by people of fairly different backgrounds. It may be useful to note the distinction between symbols and signals, as argued by Firth (1973). Firth (1973) notes that a signal is something which "tends to connote some precision of technical consequences" while a symbol connotes "a much more imprecise, open-ended sequence of events and experiences" (p. 66). It is useful to conceive of terms in categorizations schemes as symbols for the characteristics of information resources represented by said terms, rather than signals. Such conceptualization acknowledges that the cultural, situational, and individual experiential factors which color the interpretation of terms-as-symbols are many.

To return to the problematic facets: Format was a problem because its initial term set, unlike the initial term sets of other facets, was not entirely a group of like concepts. It mixed general terms, such as "Image," at the same level with more specific forms, like "Sonar Mosaic" without regard to hierarchy. For a long time, this term list remained a flat, nonhierarchical file, and when hierarchical terms, such as the term "Software" with its child term "Applet", began to be added, the original terms ("Image," "Sonar Mosaic," etc.) were not appropriately reordered to also reflect hierarchy. Additionally, the term list included some terms which were

both specific computer file types and general content descriptions (the worst of these being text, which could be interpreted either as a "Plain Text"-formatted file or as any document bearing transcribed human language rather than, say, graphics.

Eventually, through interaction with users, it became more evident that these fields, with their internal inconsistencies and their overlap of one another, were inherently confusing (they were confusing to cataloguers, as well, who could not agree on categorizations). They needed overhaul.

A more clear-cut approach to the basic shared goal of Class, Format, and Audience was found by understanding the common ground among these facets. Each of them was an attempt, however awkward, to describe how a document's author's interpret raw data, with the assumption that some types of interpretation will be more applicable to specific uses than others. A more basic, and thus more objective, way to describe how information was presented by a document would be to note its "file types" (a computer's understanding of file formats, such as "JPEG Image," which could be used to eliminate types of information that a user's computer could not process) and "content types" (a human's understanding of content formats, such as photographs of benthic fauna, bibliographies from scientific reports, etc.). Thus two new facets were developed: File Type and Content Type. Because the terms for these fields represent relatively unambiguous concepts which may be applied consistently among cataloguers, they better met the MRIB's goals of being clear to end-users and encouraging cataloguing by document authors.

A widespread, mature vocabulary for File Type already existed: the Multipurpose Internet Mail Extensions (MIME) types (ftp://ftp.isi.edu/in-notes/iana/assignments/media-types/media-types). The MRIB adopted this standard as the vocabulary for its File Type facet. Since MIME is so well-developed and used in so many applications, creating a new vocabulary for File Type would have been redundant and would needlessly complicate interoperability with other metadata standards. Content Type, on the other hand, was not a basic concept for which an adequate vocabulary existed outside of the MRIB. The nearest semblance to this facet was the Dublin Core's "Type" Vocabulary (described briefly above), which provides ten basic terms such as "Dataset" and "Events." Beginning with the DCMI "Type" term list, terms were eliminated that were not relevant to the MRIB's scope (such as "Physical Objects,"). Next, the vocabulary was expanded downwards, using the Dublin Core type terms as the upper-level categories in a hierarchical term list. Thus, interoperability with the Dublin Core metadata standard is ensured, while the vocabulary still provides more detail-rich information for use by the MRIB and MRIB-compatible systems. As of July 2003, the MRIB's EIC collections are being reworked to remove the Audience, Class, and Format facets and insert File Type and Content Type.

As the need for separate, albeit linked, item and collection metadata standards became evident, an additional facet was created to define hierarchical relationships between items and collections (or subcollections and supercollections). This facet, Collection Name, matches the field Collection Title in the Collection metadata record for the collection that includes the item. For instance, there might be a collection of satellite images of Lake Erie with an MRIB metadata record. Individual HTML documents that comprise the collection might each have their own metadata records as well. Each record for those HTML documents would then list "Lake Erie Satellite Images" as the value of their Collection Name, and the metadata record for the collection would list its Collection Title as "Lake Erie Satellite Images" to complete the link. Since the Collection Name facet is also present in Collection records, a hierarchy of

documents within a series of nested collections can be defined. For instance, that Lake Erie photo collection might be part of a larger collection of Great Lakes satellite photos, and might list "Great Lakes Satellite Images" in its own Collection Name (not Collection Title) field. Thus an interface to the MRIB metadata can guide the user from information about a collection to the individual collections or items that comprise it, or the interface may guide the user from a useful page to other pages within the same collection.

Back to Top

The MRIB Collection Metadata Facets

The MRIB metadata standard for collections is similar to that for items, except for a few additional fields (all of them possessed by collection records and absent from item records) which are listed in Figure 2. The Collection Title facet contains the title of the collection that is being indexed. The title is added to a controlled vocabulary list to ensure matching from collection records to the records of the items and/or collections that they contain. (Collection Title is not to be confused with the Collection Name facet; Collection Name defines the collection to which an item or collection belongs, whereas Collection Title is the title of a collection itself. In other words, the distinction is a way to nest collections. This may be understood by analogy to directories in a computer's file system, which may contain other directories as well as individual documents.)

The remaining collection-specific metadata fields are free-text descriptors. One of these, Collection Coverage, describes the collection's subject matter in a brief phrase. Another, Collection Alert, describes any cyclical "downtime," periodic removal of archival data, or similar information about the collection. Item Count describes the number of items within the collection, be they photos, Web pages, PDF files, or other items; this field may be omitted when the collection's nature is such that a count would be impractical or useless (for instance, the www.usgs.gov Web site can be viewed as a collection of Web pages, but it changes so frequently and is so large that a count would be neither useful nor possible). Finally, Update Frequency describes how often the collection is changed, if modifications or additions occur regularly. Ontologically, these fields are all grouped in the free-text descriptors category, except Collection Name, which belongs to the essential facets group.

Back to Top

Discussion: Meeting the Challenges of Cataloguing Digital Marine Science Resources

Earlier in this paper were listed six special challenges posed in the creation of a digital library for the marine sciences. These challenges were critical considerations during the development/expansion of the MRIB metadata fields and controlled vocabularies. The MRIB became officially "public" in January 2003 (although earlier versions were available online, they were not actively promoted). This means that it is especially timely to critically evaluate the MRIB metadata standard and the Web interface to the catalogue using that standard.

Early informal user testing of the MRIB suggests that the main user difficulties occur at the interface level. At this level, crucial considerations include 1) arranging the facets so they are all visible and clearly-purposed on the page, and 2) providing integrated definitions of words in the categorization scheme. In the meantime, despite the lack of a stable interface, it is possible to evaluate the MRIB metadata standard, at least on a preliminary basis, by considering the six challenges that we previously noted.

Back to Top

Challenge One: Accommodate geospatial and temporal "footprint" of information.

This challenge is perhaps the one most thoroughly met by the MRIB. The MRIB metadata fields include six fields specifically dealing with location (maximal, minimal, and mean latitudes and longitudes). Because these fields store information in a nearly universal format, that is, decimal coordinates, any front-end can process these data for a variety of display and matching interfaces. The current MRIB front-end uses these data to 1) plot information resources on a global map based on their area of study and to 2) match latitudinal and longitudinal ranges with entries in a gazetteer of named locations. Another location-related field, Physiographic Features, provides information about kinds of spatial features (such as mountains) rather than individual spatial features (such as Sand Mountain). Because natural processes should be very similar (or at least worth comparing) at locations which are similar feature types (for instance, two locations with coral reefs) many users will find this Physiographic Features field useful for finding relevant information about, for instance, coral reefs in general as well as in particular.

The MRIB standard also records six time-related data. The dates over which research that contributed to an information resource was carried out are noted in Research Start Time and Research Stop Time. The geological time that the information resource discusses, which for the marine sciences is often not the same as the time of the research, is placed in Geologic Time. The date when a document was last updated (prior to indexing) and the date of the last re-indexing, modification, or verification of a metadata profile are recorded in Item Last-

Updated and EIC Last Updated respectively. The date a document was first indexed using the MRIB standard is recorded in the EIC Created field.

The MRIB metadata fields for time and spatial information are very thorough. One minor problem is that, because the Geologic Time facet uses the standard geologic time scale, studies over the past 10,000 years are all grouped under either of the terms "Holocene" or "Present." Because this period is one for which a very high resolution of information is available, it would be worthwhile to divide this period into millennial- or even decadal-scale blocks.

Back to Top

Challenge Two: Integrate Information From a Broad Spectrum of Academic Disciplines

Currently, the MRIB has limited itself in scope to information from the computational, natural, and social sciences. This removes some of the difficulty involved in distinguishing scientific from artistic understanding that would be posed by incorporation of materials from the arts and humanities. The Earth and social sciences are frequently concerned with space and time, and the approach of the MRIB to such information is outlined above.

The MRIB metadata standard allows users to choose from a hierarchical list of disciplines, the Disciplines facet. Additionally, the MRIB metadata include several fields that describe information relevant to specific discipline groups. Because the biological sciences are important to oceanography, it was necessary to develop a scheme for recording information about organisms discussed in a document. The Biota facet serves this purpose by providing a Linnean hierarchy of taxonomic clades of organisms.

The Biota facet does have some drawbacks. One is that the biological taxonomic order is in constant flux, meaning that not all biologists would agree with the placement of a given organism in a given clade. Moreover, the terms currently listed in the controlled vocabulary for this facet require a scientific background; they are derived as well as possible from current scientific classification of organisms, and usually go only to the Order level of depth. It will eventually be necessary for front-ends using the MRIB metadata standard to map from scientific names to common names. Additionally, it is becoming evident the term list needs to be extended all the way to the species or subspecies level. Such detail, though inconvenient unless the cataloguer is familiar with the biologic taxonomy, will enable seamless switching between Latin organism names and folk names in any language and at any level of depth (for instance, a mapping could be developed for the vague, general term "fish" to the scientific names it includes, and it would function as well as a mapping from a scientific species name to a common species name).

Lastly, the MRIB metadata includes an Other Keywords field that can encode information not stored elsewhere. Such keywords can be used as the basis for expanding the MRIB vocabulary lists and fields as needed, as well as providing additional terms to be text-searchable in a front end (for instance, the MRIB's current Web interface enables both browsing of the facets as well as a text search that queries all of the facets and the Other Keywords field).

Back to Top

Challenge Three: Organize Information So That a Variety of Searching Strategies Can Succeed

It is often said that the imperative journalistic question is "Who, What, Where, When, Why, and How?" The MRIB metadata standard was developed with this guiding question, because users with different experiential backgrounds (and different objectives) may be most strongly appealed to by any one of these questions. Moreover, answering such questions can guide the user to browse for information along lines that seem interesting to him or her, even without a specific goal in mind. Each of the six components of the journalistic question is answered by one or more facets (and some additional information about them is stored in non-searchable metadata). "Who?" is answered by Authors, Agencies, and Projects. "What?" is answered by Disciplines, Features, and Biota. "Where?" is answered by Location. "When?" is answered by Geologic Time. "Why?" is answered by Hot Topics, and "How?" is answered by Methods, Content Type, and File Type.

Moreover, the vocabularies are designed to address concepts redundantly by alluding to a single concept in different facets, with each occurrence of the concept being tempered by its relationship to the whole facet. For instance, a researcher interested in sediments might find relevant resources through the avenues noted in Table 2.

Table 2: Subconcepts of the “sediment” concept.

TERM	FACET
Geology/Sedimentology	Disciplines
Geochemistry/Sediment Geochemistry	Disciplines
Soil	Physiographic Features
Geological Features/Sediments	Physiographic Features
Environment/Environmental Issues Relating to Sediments	Hot Topics
Environment/Environmental Issues Relating to_ Habitats/Sediments in Habitats	Hot Topics
Disasters/Types of Disasters/Erosion	Hot Topics
Disasters/Types of Disasters/ Subsidence	Hot Topics
Field Observation Methods/Sampling Methods/Surface Sampling Methods	Methods
Field Observation Methods/Sampling Methods/Sampling Methods Using Cores	Methods

The terms and facets shown in Table 2 all represent subtle aspects of a single

concept, "sediment." It should be noted that this redundancy does not mean the same concept is represented by different terms (which would defeat the purpose of a controlled vocabulary); rather, many variations or sub-concepts of a broader concept are represented. A sophisticated front-end to the MRIB metadata standard might analyze these related concepts and provide "Related To..." options for the user.

Back to Top

Challenge Four: Minimize Jargon

Generally, the MRIB standards minimize argot. However, in some cases, it becomes inevitable. The Biota facet is an example of this: there is no precise way to describe organisms besides the taxonomic standard (a "standard" which is really in flux). Nonetheless, the taxonomic standard was chosen with the intent that it could be hidden from non-biologist users under a variety of interfaces that would be dependent on the precise nature of the scientific taxonomy. Another such case is Geologic Time, which again relies on standard naming conventions with which some users may be unfamiliar. In these instances, it is necessary for the front-end to provide term definitions and guidance to the user.

This is where the additional information stored in the MRIB's valid term lists becomes handy. These lists, in tab-delimited form, store not only term names but also definitions of the terms that can be incorporated into a front-end that reads the MRIB standard. (These standardized definitions also prove useful to indexers, because they may provide more precise connotations than a term generally carries, or clarify terms that are differently applied among academic disciplines.)

Where possible, the MRIB avoided terms that were likely to be confusing. For instance, terms with multiple meanings across fields were avoided. One example of this was the choice to use "soil science" for the geological field of "pedology" in the Disciplines facet because it sounded too similar to the field of "pediatrics", the medical treatment of children.

Back to Top

Challenge Five: Use Enough — But Not Too Many — Metadata Fields

When adding new fields, the MRIB team was cautious, and verified that concepts could not be incorporated logically into existing fields to avoid the fission of essentially similar concepts into an infinitely large (not to mention confusing) set of browseable fields. In some cases, facets were actually merged when unexpectedly significant overlap with other facets became evident (such as the Format, Audience, and Class facets, which were transformed into File Type and Content Type).

Despite this cautious approach to creating facets, the number of faceted metadata fields is still large enough to pose an interface design hurdle. Making each of the facets available, and their potential usefulness clear, is an ongoing process. As the interface is refined through continual adjustment in response to user testing and feedback, we will be able to discern with some clarity whether the MRIB fields are few enough to function well in an entry-level user interface. Regardless, it is possible to develop multiple user interfaces, some of which present for browsing

the full breadth of metadata information available, and others which simplify the scheme to meet less rigorous (or more specific) searching needs.

Back to Top

Challenge Six: Encourage Composition of Metadata Records by Resource Creators Themselves

This challenge is closely tied to the other challenges, and is also a broad test of whether the categorization scheme is intuitive and consistent. If different cataloguers who have not extensively used the MRIB metadata standard can develop very similar records for the same document, that is solid evidence in favor of intuitiveness and consistency. (The records need not be exactly the same, since some of the free-text fields, such as abstract, will inherently vary.) Because the cataloguing process is independent of the main MRIB interface, and thus freed of interface concerns, feedback from document authors and maintainers who have catalogued their own Web pages may provide more insight into the sturdiness of the categorization scheme than user testing of the main MRIB. (Although a cataloguing interface is involved if the cataloguer chooses not to generate records manually, this interface proceeds through each metadata field sequentially, unlike that of the main Web interface, so it is ensured that users see each field and have the opportunity to use it.)

So far, cataloguing by people other than the MRIB staff has been successful so long as cataloguers are provided with concise definitions of terms. When definitions are absent, these cataloguers often become overwhelmed with the sheer number of metadata fields. However, with the term definitions in hand, the users are usually able to compose records that agree with those created by the MRIB staff.

Other elements of meeting this sixth challenge involve providing a cataloguing interface (and promoting it) to document authors who do not wish to enter the messy-looking world of manually encoding catalogue data in the EIC format. However, this is a relatively simple matter compared to providing an end-user interface to the metadata records, so it will not be elaborated here.

Back to Top

Conclusion

As scientific information has been increasingly made available on the Web, specialized means to organize and integrate such information have become necessary. Some attempts to meet this need are very general in purpose, while others, like the MRIB standard, have been refined for specific audiences, purposes, and scopes. Because this standard is intended for a broad audience and has a narrow but heterogeneous scope (information from the marine, lake, and coastal sciences, which draw from a variety of disciplines), its development provides useful lessons for the creation of other distributed libraries.

The development of an organic categorization scheme directly from the catalogued documents, and using terms from the documents themselves, was an expedient way to produce detailed controlled vocabularies. Mis-steps along the way largely involved ambiguous terms and overlapping facets. By striving to clarify terms and promote homogeneity of terms within facets, some of these problems have been removed.

The MRIB metadata development was also a lesson in how an ontology should, and should not, be constructed. The "indexing-to-discover" process proved very useful in expanding the term list, but the failure of the MRIB developers to respond early to apparent structural problems (such as in the example of the Class, Audience, and Format facets) has required time-consuming revisions of the metadata records, revisions of a sort that will not be possible once the MRIB categorization scheme begins to be used widely. Moreover, it seems that the early emphasis on geology left the preliminary categorization scheme lacking fields to describe important geographical and biological concepts. Perhaps earlier collaboration with non-geologist marine scientists would have made the need for the Physiographic Features and Biota fields more evident from the beginning (and would have provided more insight into how Biota might be best structured). In short, a process of indexing to discover terms, combined with a willingness to make structural revisions at the early stages, seems to be an ideal approach to developing a categorization scheme of this nature.

At this stage, the MRIB scheme has been used to index thousands of electronic resources. Further development of the scheme will likely focus on developing a supportive front end that will ensure that terms are clear to most end-users. To this extent, describing the challenge in anticipating users' search strategies, Bates (1998) stated:

> ...The better developed the typical system, the more arcane its fine distinctions and rules are likely to be, and the less likely to match the unconsidered, inchoate attempts of the average user to find material of interest. [The] question should not be: 'How can we produce the most elegant, rigorous, complete system of indexing or classification?,' but rather, 'How can we produce a system whose front-end feels natural to and compatible with the searcher, and which, by whatever infinitely clever internal means

we devise, helps the searcher find his or her way to the desired information?'

With a solid EIC creation system now available to its would-be cataloguers, the MRIB faces the challenge of building support structures that will guide the user, through a variety of means, to useful information. This support structure will include the availability of definitions of terms, and will likely include searchable indices of related words linked to the MRIB's controlled terms. Other useful infrastructure is already available to the MRIB; the visibly faceted categorization scheme allows the user to choose a facet, then to see how one item from that facet intersects with the matches for a term from another facet. This capacity for guided wandering allows the user to both zoom in and pull back at his or her choosing.

Although the MRIB metadata were intended to work with the MRIB Web interface, the standard is open, and full information about its application is available on the MRIB Web site. This openness encourages the use of MRIB metadata fields, terms, and even metadata records by other applications. It is possible, for instance, to envision an implementation of the MRIB metadata that would regularly "spider" the Web seeking MRIB metadata stored within XML files or HTML tags, allowing authors to more readily update metadata for their documents. It would also be possible to adapt the MRIB metadata to terrestrial Earth science information, with the addition of methodological terms and "hot topics" more applicable to the continental realms. The flexible, hierarchical structure of the MRIB categorization scheme permits the development of MRIB-based systems that look radically different from one another. An interface to the categorization scheme might display only a subset of the available facets and terms, or truncate them at any point, to provide a simpler interface. Or the interface might have its own set of terms, each of them linked (invisibly to the end-user) to either one or more of the terms in the underlying ontology (perhaps in a different language, dialect, or technical level). An adventurous system might add its own, deeper levels to the terms.

These deeper levels could be used by its specialized interface, and ignored by other systems whose audiences had no need for such levels (or adopted by other systems which did need the specialized levels).

Ultimately, the MRIB is an ongoing project, and doubtless it will evolve much over the next few years. However, because the MRIB metadata represents a stable means for the classification of information resources about Earth's water bodies, the MRIB metadata may serve as a foundation on which a variety of user interfaces and metadata records can be built.

Back to Top

Title / Introduction / Cataloguing / Evolution / Case Study / Discussion / Conclusion / References /

Department of Interior / **U.S. Geological Survey** / **Coastal and Marine Geology**

USGS Privacy Statement / Disclaimer / Accessibility

This is http://woodshole.er.usgs.gov/drafts/accessib/meta_standards/html/concl.html
Maintained by webmaster-woodshole@usgs.gov
Modified Tuesday, 02-Mar-2004 11:20:22 Eastern Standard Time

References

Bates, M. J., 1998, Indexing and Access for Digital Libraries and the Internet, in Human, Database, and Domain Factors: Journal of the American Society for Information Science, v. 49 (November 1998), p. 1185-1205.

Dublin Core Metadata Initiative, 2003a, DCMI Frequently Asked Questions (FAQ): Dublin Core Metadata Initiative Home Page. **ONLINE** .

Dublin Core Metadata Initiative, 2003b, DCMI Type Vocabulary; Dublin Core Metadata Initiative Usage Board: Dublin Core Metadata Initiative Home Page. **ONLINE** . (Version 12FEB2003).

Federal Geographic Data Committee, 2000, Content Standard for Digital Geospatial Metadata Workbook: Federal Geographic Data Committee, Reston, Virg., FGDC Home Page. **ONLINE** . (Version 31MAY2000). (PDF format)

Firth, Raymond, 1973, Symbols Public and Private: London: George Allen & Unwin, 1973.

Furrie, B., 2000, Understanding MARC Bibliographic, Machine-Readable Cataloguing: Library of Congress Network Development and MARC Standards Office, Washington, DC, MARC Home Page. **ONLINE** . (Version 9JAN2003).

Gilliland-Swetland, A. J., 1998, Defining Metadata, in M. Baca, (ed.) 1998, Introduction to Metadata: Pathways to Digital Information, Getty Information Institute, Los Angeles, USA, 41 p.

Lerner, S. and Maffei, A. (2001), 4DGeoBrowser, A Web-Based Data Browser and Server for Accessing and Analyzing Multi-Disciplinary Data: Technical Report WHOI-2001-13, Woods Hole, Woods Hole Oceanographic Institution, 2001.

Library of Congress Network Development and MARC Standards Office, Washington, DC: MARC Home Page. **ONLINE** . (Version 9JAN2003).

Library of Congress, 2002, Guidelines for Coding Electronic Resources in Leader/06: Library of Congress Network Development and MARC Standards Office, Washington, DC, MARC Home Page. **ONLINE** . (Version 31DEC2002).

Marincioni F., Lightsom F. L., Riall, R.L., Linck, G.A., Aldrich, T.C., Caruso M.J., in press, The marine realms information bank, a distributed geolibrary for coastal and marine science: Journal of Digital Libraries, 2003.

Miller, D. R., 2000, XML and MARC, A Choice or Replacement?: American Library Association MARBI/CC, DA Joint Meeting, Chicago, Lane Medical Library

Title / Introduction / Cataloguing / Evolution / Case Study / Discussion / Conclusion / References /

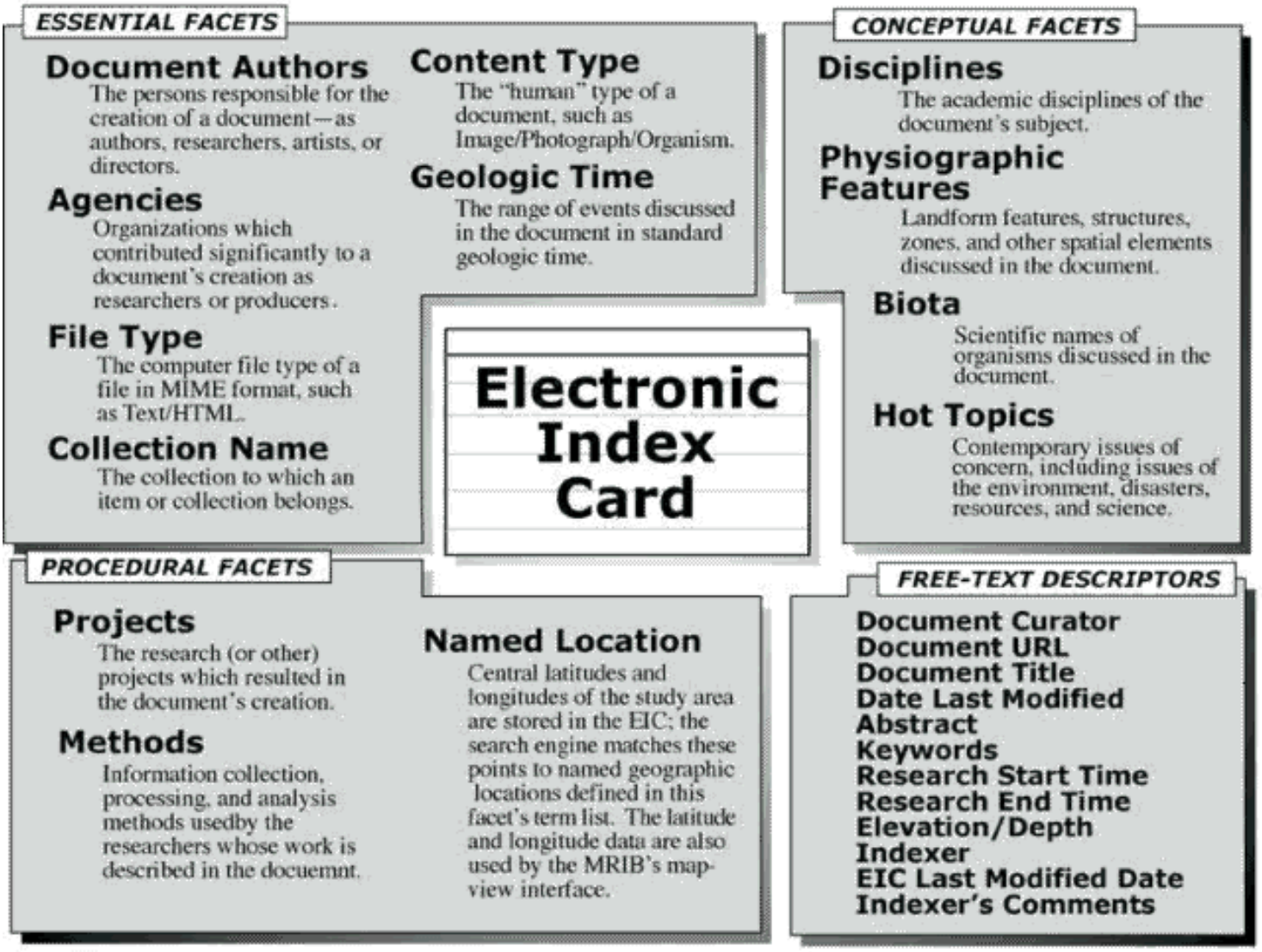
Department of Interior / **U.S. Geological Survey** / **Coastal and Marine Geology**

USGS Privacy Statement / Disclaimer / Accessibility

This is http://woodshole.er.usgs.gov/drafts/accessib/meta_standards/html/ref.html
Maintained by webmaster-woodshole@usgs.gov
Modified Friday, 06-Feb-2004 08:29:54 Eastern Standard Time

Figure 1.

The MRIB's item metadata fields.



Department of Interior / **U.S. Geological Survey** / **Coastal and Marine Geology**

This is http://woodshole.er.usgs.gov/drafts/accessib/meta_standards/html/fig_pages/fig1.html
Maintained by webmaster-woodshole@usgs.gov
Modified Thursday, 29-Jan-2004 13:22:16 Eastern Standard Time

---

Figure 2.

The additional metadata fields for collections.

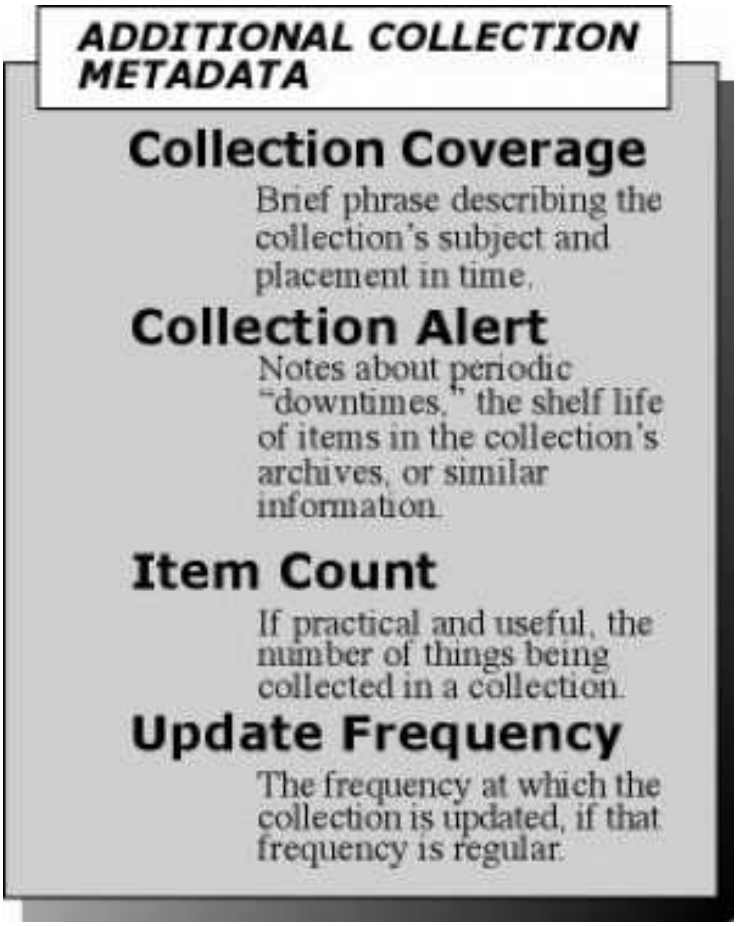


Figure 3.

Item and collection "nesting."

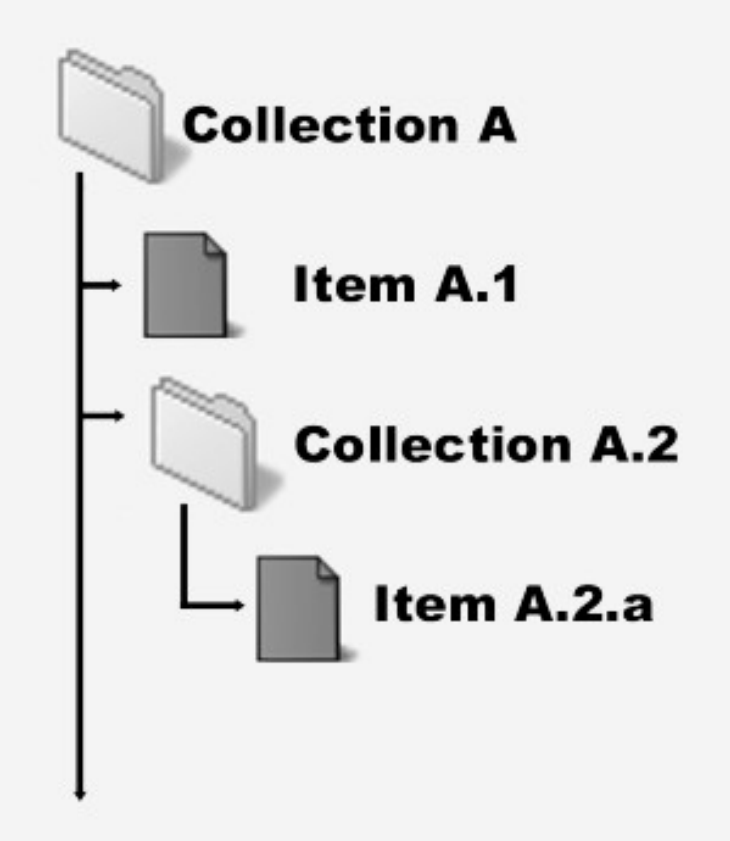


Figure 4.

Controlled vocabulary for one facet, Physiographic Features, in action.

Geographical Features → 223 matches

- Benthos → 3 matches
- National Marine Sanctuary → 221 matches
- Archaeological Site → 34 matches
- State Park → 2 matches
- Superfund Site → 1 match

Geological Features → 59 matches

- Accretionary Wedge → 26 matches
- Fault → 221 matches
- Formation → 1 match
- Magma Chamber → 2 matches
- Ophiolite → 3 match
- Plate → 5 matches
- Plate Margin → 2 matches
- Archaeological Site → 34 matches
- Rift → 5 matches
- Rift Zone → 3 match
- Sediment → 18 matches
- Spreading Center → 7 matches
- Subduction Zone → 1 matches
- Spreading Center → 7 matches

Biological Features → 347 matches

- Bench Zones → 7 matches
- Benthos → 92 matches
- Coral Reef → 1 match
- Forest → 3 matches
- Kelp Forest → 9 match
- Marsh → 18 matches
- Peat Bog → 1 matches
- Prairie → 1 matches

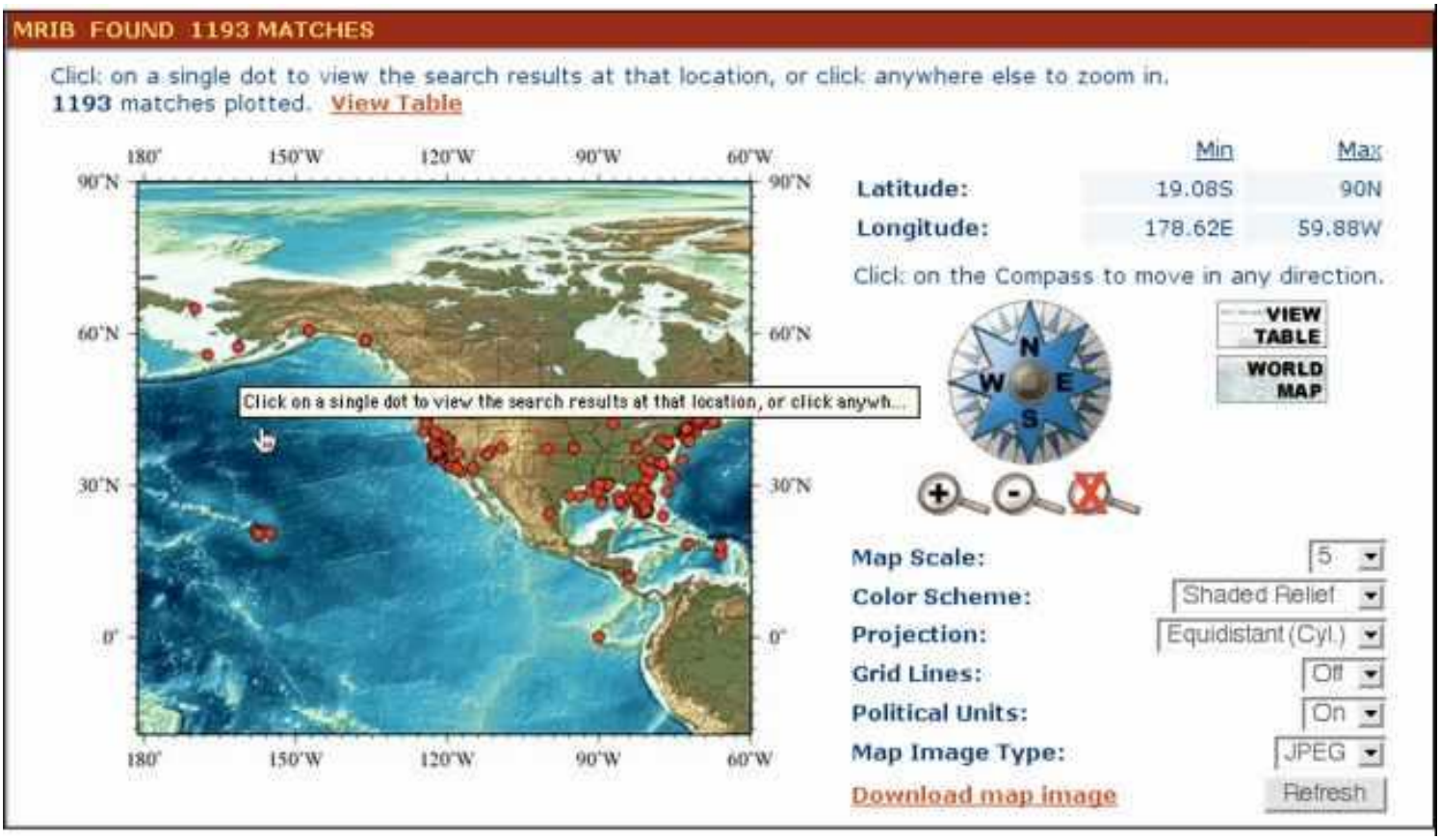
- Tundra → 14 matches
- Water Column → 262 match

Title / Introduction / Cataloguing / Evolution / Case Study / Discussion / Conclusion / References /

Department of Interior / **U.S. Geological Survey** / **Coastal and Marine Geology**

Figure 5.

Map view and named locations are two ways the MRIB interface provides the spatial context of information resources.



Above: Locations represented visually as points on a map.
Below: Locations represented textually as hierarchically-arranged named areas.

1193 Matches. View Entire Tree, Table or Map of all matches.

To search a sub-category, click on ther term link, next to the list bullet. To search and display the table of search results, click on the link to the right of the term (ie, --> 15 matches).

Oceans | Continental Coasts | Seas and Gulfs | Lakes | Geopolitical Units

Location:

- Oceans
  - Artic Ocean --> 1 match
  - Atlantic Ocean --> 727 match
  - Pacific Ocean --> 1193 match
- Continental Coasts
  - Americas --> 1193 match
- Seas and Gulfs
  - Americas --> 1193 match
  - Asia --> 4match
- Lakes
  - North America --> 1137 match
  - Caribbean --> 134 match
  - Western Pacific --> 23 match
  - South America --> 55 match
- Geopolitical Units
  - Americas --> 1193 match
  - Caribbean --> 134 match
  - Central America --> 55 match
  - Central Pacific --> 23 match
  - Western Pacific and Australia --> 23 match

Title / Introduction / Cataloguing / Evolution / Case Study / Discussion / Conclusion / References /

Department of Interior / **U.S. Geological Survey** / **Coastal and Marine Geology**

USGS Privacy Statement / Disclaimer / Accessibility

This is http://woodshole.er.usgs.gov/drafts/accessib/meta_standards/html/fig_pages/fig5.html
Maintained by webmaster-woodshole@usgs.gov
Modified Tuesday, 03-Feb-2004 09:41:12 Eastern Standard Time

Figure 6.

Term lists from the old Format, Class, and Audience facets.

AUDIENCE:	**FORMAT**:	**CLASS**:
Disaster_Management	Atlas	Data
Model_Implementation	Data_Set	Derived_Products
Policy_Making	Data_Set.DODS	Knowledge
Public_Awareness	Database	Predictions
Recreation	Dynamic_3d	Scientific_Publications
Resource_Management	Email_List	
Teaching	Equation	
Equipment_Use	FTP	
	GIS	
	Graph	
	Image	
	Map	
	Mosaic	
	Movie	
	Newsgroup	
	Photograph	
	Poster	
	Software	

Software.Applet

Software.Extension

Software.Source_Code

Sound

Tabular

Text

Title / Introduction / Cataloguing / Evolution / Case Study / Discussion / Conclusion / References /

Department of Interior / **U.S. Geological Survey** / **Coastal and Marine Geology**
