Chapter XLIII
USGS Digital Libraries for Coastal and Marine Science

Frances L. Lightsom
U.S. Geological Survey, USGS Woods Hole Science Center, USA

Alan O. Allwardt
ETI Professionals, USGS Pacific Science Center, USA

ABSTRACT

The U.S. Geological Survey (USGS) has developed three related digital libraries providing access to topical and georeferenced information for coastal and marine science: the Marine Realms Information Bank (MRIB) and its two offshoots, the Monterey Bay Science Digital Library and Coastal Change Hazards Digital Library. These three members of the MRIB family run on the same software and share a common database, but they employ different user interfaces targeting different audiences. This chapter reviews (1) distributed geolibraries, the conceptual foundation for MRIB, (2) the modular software of MRIB, permitting the rapid development of customized user interfaces, and (3) the Electronic Index Card (EIC) Creation Utility, encouraging users to contribute new metadata records to the MRIB database. The accompanying discussion addresses several challenges facing digital library developers: providing for scalability in the system; ensuring interoperability with other systems; and meeting the demands of characterizing information while facilitating its search and retrieval.

INTRODUCTION

The Coastal and Marine Geology Program (CMGP) of the U.S. Geological Survey (USGS) has developed a family of distributed digital libraries providing access to topical and georeferenced information for coastal and marine science. This digital library system includes three user interfaces targeting different audiences:

1. The Marine Realms Information Bank (MRIB; http://mrib.usgs.gov/), developed in 2001, is a general-purpose user interface providing access to free online scientific in-

formation about oceans, coasts, and coastal watersheds. MRIB encourages its users to discover these information resources by browsing a faceted classification with twelve main categories, including author, agency, discipline, feature type, named location, and "hot topics" (Figure 1). MRIB was also one of the first digital libraries to utilize interactive maps for searching and retrieving georeferenced information.

2. The geographic search capabilities of MRIB were ideally suited for creating the Monterey Bay Science (MBS) Digital Library (http://mrib.usgs.gov/mbs/), a regional pilot project providing access to scientific information about the Monterey Bay National Marine Sanctuary and coastal watersheds of central California (Figure 2). The MBS user interface, released in 2004, serves as a model for any regionally focused digital library based on the MRIB software architecture.
3. The newest addition to the CMGP digital library system is the Coastal Change Hazards (CCH) Digital Library (http://mrib.usgs.gov/cch/), released in 2006. The specialized CCH user interface (Figure 3) focuses on natural hazards and human impacts in the coastal zone and replaces the MRIB hot topics with a more specific topical classification. Crosswalks between the MRIB and CCH topical classifications ensure that online resources originally cataloged for one interface can be searched and retrieved using the other interface. The Coastal Change

Figure 1. The Marine Realms Information Bank (MRIB), featuring three search options: by category, location, and keyword. The "Submit a Document" link at the bottom of the page connects the user to the Electronic Index Card (EIC) Creation Utility

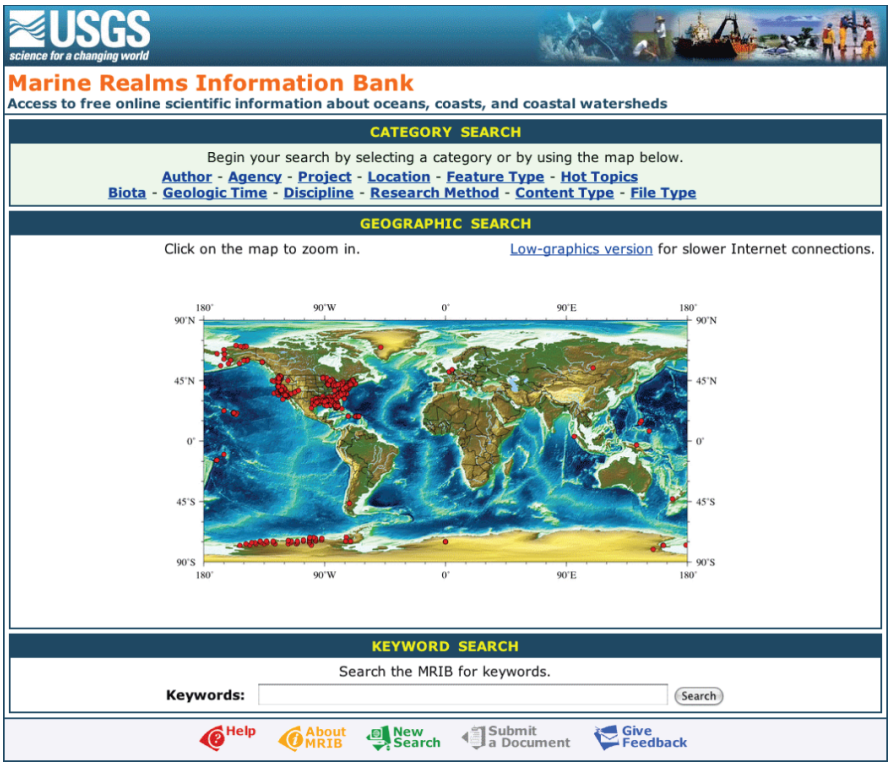


Figure 2. The Monterey Bay Science (MBS) Digital Library, a regionally focused member of the Marine Realms Information Bank (MRIB) family. The customized MBS user interface provides access to about one-fourth of the MRIB database

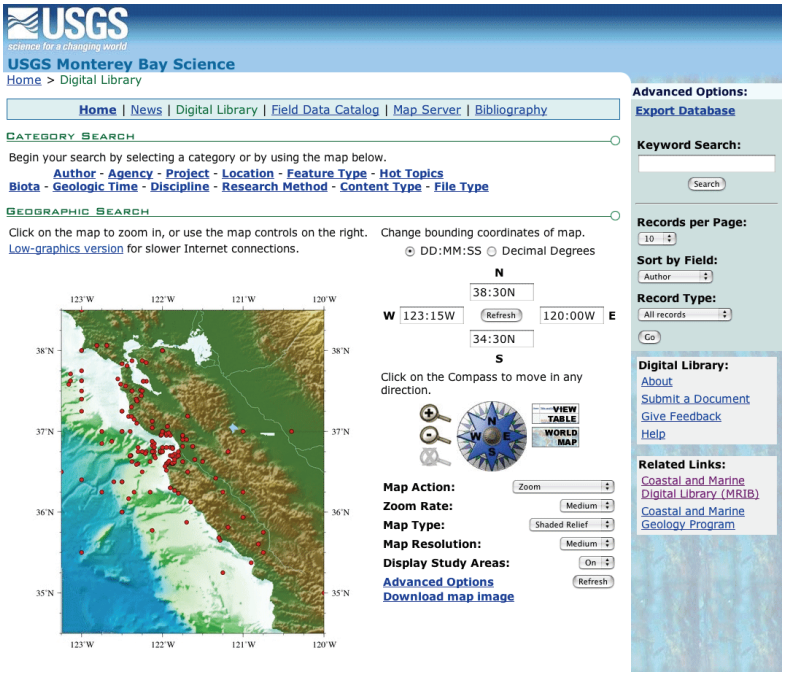
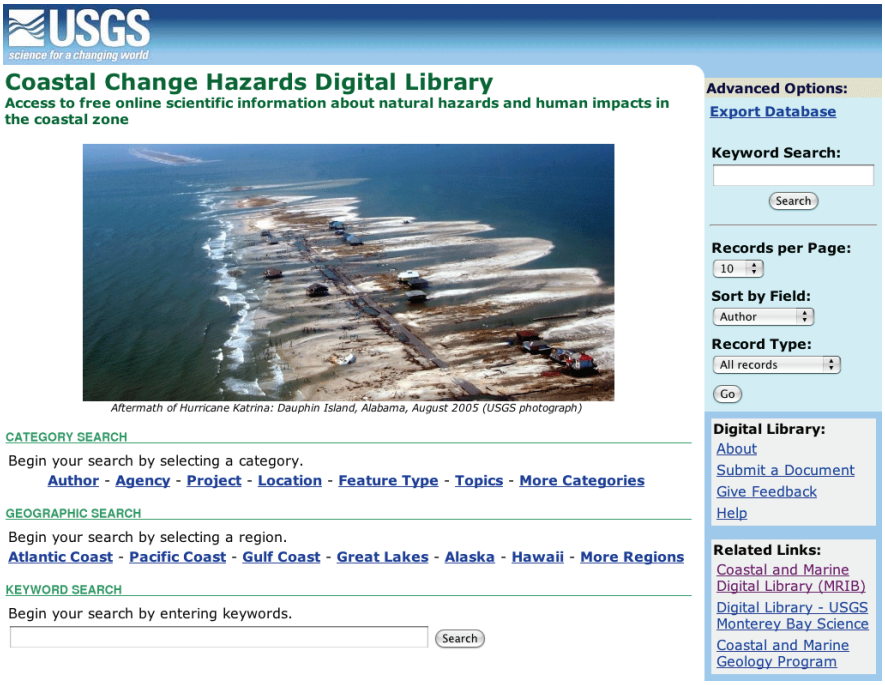


Figure 3. The Coastal Change Hazards (CCH) Digital Library, a topically focused member of the Marine Realms Information Bank (MRIB) family. The customized CCH user interface provides access to about one-third of the MRIB database



Hazards Digital Library serves as a model for any topically focused digital library based on the MRIB software architecture.

These three closely related digital libraries, which run on the same software and share a common database, constitute the MRIB family of digital libraries. In the discussion that follows, the term "MRIB" will be used in a generic sense for all three members of the family, whereas "Marine Realms Information Bank" will refer specifically to the parent interface.

BACKGROUND

The rationale for MRIB can be found in two National Research Council (NRC) studies released in 1999. One study, *Science for Decisionmaking*, was an external review of the USGS Coastal and Marine Geology Program (CMGP) conducted by NRC at the request of the USGS. The final report posed three "grand challenges" for CMGP over the next few decades, including development of a "national knowledge bank on the geologic framework of the country's coastal and marine regions" (Committee to Review, 1999, p. 48–49). In response to this NRC recommendation, CMGP has designed a knowledge bank prototype including three complementary components: a digital library (Marine Realms Information Bank), a field data catalog (InfoBank), and an Internet map server/GIS data catalog (see http://marine.usgs.gov/kb/).

The other NRC study, titled *Distributed Geolibraries*, was characterized by its authors as a "vision for the future" of information retrieval: providing access to online resources in response to geographic queries (Panel on Distributed Geolibraries, 1999). The Alexandria Digital Library, a collaborative project coordinated by the University of California, Santa Barbara, was an early test bed for this concept, addressing fundamental issues in the search and retrieval of georeferenced information resources (see Janée, Frew, & Hill, 2004). Another successful Web-based system for organizing and accessing georeferenced information is the 4DGeoBrowser, developed at the Woods Hole Oceanographic Institution (Lerner & Maffei, 2001). The flexible 4DGeoBrowser software has been used to create more than a dozen separate Web applications serving the oceanographic community, including the initial version of the Marine Realms Information Bank (see http://4dgeo.whoi.edu/ for links to these individual applications). Related research on the design of coastal Web atlases is summarized in O'Dea, Cummins, Wright, Dwyer, and Ameztoy (2007), with links to several examples.

MRIB DESIGN

MRIB is a distributed geolibrary providing access to selected online information resources for coastal and marine science, including Web sites, full-text reports, digital maps, and downloadable data. MRIB does not store original data or information on its server but rather the metadata records and hyperlinks for online resources maintained on other servers. The MRIB user interface facilitates searching by topical category, location, or keyword, and the metadata records can be downloaded in several formats (including plain text, comma-separated values, XML, and KML).

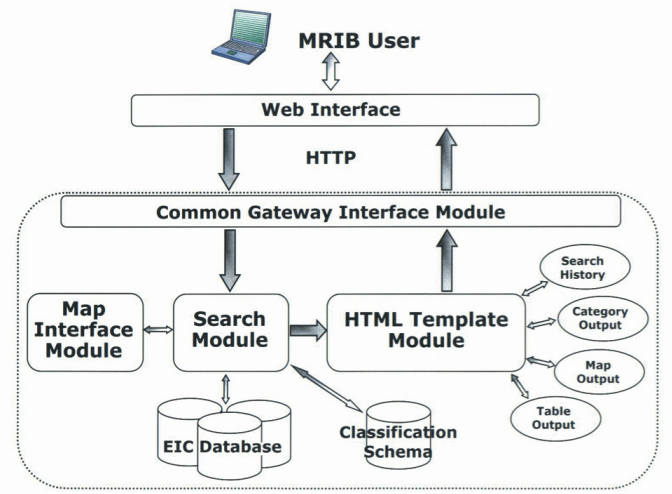
MRIB employs a faceted classification for topical searching. Each of the twelve main categories includes a controlled vocabulary representing a hierarchy of associations or concepts. This knowledge organization system (KOS) is designed to accommodate a wide range of users, including scientists, public servants, advocacy groups, educators, students, and the general public (for additional details on the KOS employed in MRIB, see Marincioni, Lightsom, Riall, Linck, Aldrich, & Caruso, 2004; for a general discussion of KOS applications in digital libraries, see Hodge, 2000).

Most of the online resources cataloged in MRIB are georeferenced and thus searchable by map coordinates or place name. The MRIB gazetteer currently lists more than 1,700 place names along with their rectangular bounding boxes (see http://mrib.usgs.gov/meta/location.txt). Subcategories in the gazetteer include oceans, continents, geopolitical units, administrative units (exclusive economic zones, marine sanctuaries, and hydrologic regions), and marine cadastres.

The three basic search operations—topical, geographic, and keyword—can be applied repeatedly, in any combination, until the search parameters are appropriately focused. At each stage of this process MRIB displays the search results in browsable tables or interactive maps (at the user's discretion) and provides links to the original online resources. The map view allows users to display search results in a variety of spatial contexts by selecting the desired background map layers: latitude and longitude grid lines, political units, exclusive economic zones, marine cadastres, and coastal hydrologic regions. Experience has suggested that a "lightweight" GIS functionality of this type is ideal for georeferenced digital libraries with a varied clientele (see Janée et al., 2004).

The modular software architecture of MRIB (Figure 4) has permitted rapid development of two customized user interfaces (in addition to the original interface): the regionally focused Monterey Bay Science (MBS) Digital Library and the topically focused Coastal Change Hazards (CCH) Digital Library. Crosswalks between the generalized topical classification employed by the Marine Realms Information Bank (MRIB hot topics) and the specialized topical classification employed by the Coastal Change Hazards Digital Library (CCH topics) ensure full interoperability between these interfaces. In the MRIB topical hierarchy there are four top-level terms: Environment; Hazards and Disasters; Resources; and Science and Scientists. In contrast, the CCH topical hierarchy has three, very different top-level terms: Agents of Coastal Change; Effects of Coastal Change; and Human Responses to Coastal Change. Nevertheless, many *specific* topics can be found at the lower levels of both hierarchies: Hurricanes and Typhoons; Tsunamis; Climate Change; Saltwater Intrusion; and so on. Constructing the crosswalks, therefore, was simply a matter of mapping semantically equivalent "pigeonholes" from one topical hierarchy to the other. In operation, this process is

Figure 4. MRIB software architecture, illustrating the modular design that permits the rapid development of customized user interfaces like the Monterey Bay Science Digital Library and the Coastal Change Hazards Digital Library (see figures 2 and 3)



transparent to the user and allows a single MRIB database to be searched from two different topical perspectives.

All digital libraries that provide access to information on other servers will be plagued, sooner or later, by dead links: much of the information found online is, by nature, ephemeral (Nelson & Allen, 2002). To combat this problem MRIB has incorporated software to detect standard 404 error messages ("Not Found"), but this measure solves only part of the problem. Many Webmasters employ "soft 404" messages, whereby invalid Web addresses redirect to a valid Web page offering a customized error message ("We're sorry, but the page you requested cannot be found …"). Alternatively, invalid addresses for lower-level Web pages might simply redirect to the home page; Webmasters often resort to this tactic when they overhaul the URL syntax for their Web sites, in order to help long-time users adjust to the sudden change. While "soft 404" messages may have merits (tips for re-navigating are usually offered), they do complicate the task of writing programs to detect dead links automatically. For this reason, the MRIB cataloger still relies upon labor-intensive inspection (and blind luck) to find and remove many dead links. Occasionally a dead link can even be resurrected (in a sense) by finding a cached version of the Web page in the Internet Archive (http://www.archive.org/). The MRIB cataloger has employed this strategy to re-establish access to a few online documents with enduring value.

Although the MRIB Team compiled the initial catalog and will continue adding entries, the creators of scientific information can be uniquely qualified to catalog their own online resources (especially in complex or newly emerging fields of research). Consequently, MRIB encourages its users to submit new metadata records with the Electronic Index Card (EIC) Creation Utility, a series of online forms and menus for the numerous controlled-vocabulary and free-text metadata fields. Users can open personal MRIB accounts, with password-protected files and directories for arranging the provisional EICs they have submitted. These accounts allow user-contributors to preview, rearrange, edit, or delete their own cards prior to final approval by the MRIB Team (as well as update cards *after* approval). The MRIB Team reviews all user contributions for suitability and accuracy before incorporating them into the public database. As this database grows, user participation will become even more important for updating old metadata records and weeding out dead links that the MRIB software is unable to detect.

Software Specifications

MRIB has been created with open-source software and open standards in the public domain. The code base of MRIB is written in Perl, version 5.8.8 (http://www.perl.com/). Generic Mapping Tools (GMT), version 4.1.1 (http://gmt.soest.hawaii.edu/), with modifications by the MRIB programmer, Guthrie Linck, is used to generate the base maps and data plots. Elevation data from several sources are used in conjunction with GMT: the ETOPO1 one-minute global relief database (http://www.ngdc.noaa.gov/mgg/global/global.html); the SRTM 1- and 3-arc-second land data (http://srtm.usgs.gov/); and the NGDC 3-arc-second coastal relief model for the conterminous United States, Hawaii, and Puerto Rico (http://www.ngdc.noaa.gov/mgg/coastal/coastal.html). Map images are available in JPEG (the default), GIF, and PNG formats. The MRIB public Webserver is a dual 2.4 GHz processor running the Apache HyperText Transfer Protocol (HTTP) server, version 2.2.3 (http://www.apache.org/), along with the Apache perl module (mod_perl, version 2.0.2). The operating system is Ubuntu Linux 8.04 (http://www.ubuntu.com/).

FUTURE DIRECTIONS

Bates (1998, 2002) discusses digital library design in the context of fundamental research by information scientists over a period of several decades. One important design consideration is the *scalability* of the system, an issue that can manifest itself in subtle ways. For instance, will a controlled vocabulary initially designed for a small digital library function as intended in a much larger system? Will a particular search strategy remain effective when the digital library grows? Along these lines, Bates argues that domain size, both current and projected, should be taken into consideration when designing search services for a digital library: browsing is feasible only in a small domain; directed searching becomes necessary as the domain increases in size; and linking is especially effective in a very large domain (e.g., the Internet).

MRIB allows its users to employ browsing, directed searching, and linking at different stages of the search process. The MRIB database is large enough to require directed searching (topical, geographic, or keyword) of its metadata records in order to retrieve a manageable subset that can be easily browsed. One form of linking is obvious, of course—when the user visits an external information resource for which MRIB provides a metadata record. Viewed as a *search mechanism*, however, linking is indirect (and serendipitous) in a distributed digital library like MRIB: the external resources themselves may provide links to additional online information that has not been indexed in the library.

Browsing serves another function in MRIB, and here the need for additional fine-tuning has become apparent. The faceted classification includes twelve categories, each with a detailed controlled vocabulary designed for precise, domain-specific resource description. By browsing these vocabulary lists the user selects as many topical subcategories as are needed to narrow the search. The underlying principle is simple, as noted by Bates (1998): most users can recognize the information they need more readily than they can recall it. As the MRIB database has grown, however, the original goal of creating precise, domain-specific resource descriptions has had an unintended consequence: the controlled vocabularies for most of the twelve facets have become so long and complex that *browsing* them raises serious usability issues for both the indexer and the user. Thus, the decision has been made to simplify these controlled vocabularies (with some obvious exceptions, such as the author list).

These changes in the MRIB metadata scheme will also (1) improve interoperability with the more generalized topical categories of the USGS Thesaurus (http://www.usgs.gov/science/) and (2) facilitate harvesting through a federated service such as the Open Archives Initiative (see Lagoze & Van de Sompel, 2001). Achieving these goals will require striking the proper balance between the finely granular metadata required for domain-specific resource description and the coarsely granular metadata appropriate for cross-domain search and retrieval (Lagoze, 2001).

To improve geographic searching, the rectangular latitude-longitude bounding boxes in the MRIB gazetteer will be soon replaced with polygonal footprints in order to eliminate many of the false drops inherent in the current system. Most of North America, for example, falls within the bounding box for the Pacific Ocean. Consequently, a gazetteer search for the Pacific Ocean yields more than 50% false drops because of the numerous hits for the Atlantic and Gulf coasts of North America. This example may be extreme because of the peculiar shape of the Pacific basin, but it serves to illustrate a general problem for any georeferencing system relying upon bounding boxes that are *necessary but not sufficient* to define irregular geographic areas.

With appropriate modifications, MRIB software could accommodate geospatial information from a wide range of natural and social sciences. The MRIB content metadata standard and soft-

ware protocols will be documented in a forthcoming USGS publication to facilitate adapting the MRIB system for other disciplines.

CONCLUSION

Lagoze, Krafft, Payette, and Jesuroga (2005) suggest that digital libraries differ from Internet search engines by *adding value* to online resources. A well-designed digital library offers a carefully selected, manageable collection of resources, organized to provide context and to highlight interrelationships that might otherwise be overlooked. Lagoze and his colleagues also argue that a digital library should encourage user collaboration in order to benefit from the "wisdom of crowds." The MRIB family of distributed digital libraries illustrates these points. MRIB adds value to online resources for coastal and marine science by providing selectivity, topical context, and spatial context. The Electronic Index Card (EIC) Creation Utility allows users to contribute new metadata records and help the MRIB Team keep the old ones current. This "grassroots" collaboration by users will also ensure that MRIB remains abreast of emerging research.

As a public science agency, the USGS is responsible for delivering timely, reliable data and information essential to meeting national needs and international obligations (Committee on Future Roles, 2001; Hutchinson, Sanders, & Faust, 2003). In the marine realm, the role of the USGS as an information agency has taken on added importance as the United States moves toward an integrated ocean policy (see U.S. Commission on Ocean Policy, 2004). The USGS Coastal and Marine Geology Program (CMGP) fulfills this responsibility in part by creating digital libraries like MRIB for a wide range of users wishing to learn about coastal and marine science, gather data for research, and make informed decisions.

DISCLAIMER

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

REFERENCES

Bates, M.J. (1998). Indexing and access for digital libraries and the Internet: Human, database, and domain factors. *Journal of the American Society for Information Science, 49,* (13), 1185–1205.

Bates, M.J. (2002). Speculations on browsing, directed searching, and linking in relation to the Bradford Distribution. In H. Bruce, R. Fidel, P. Ingwersen, & P. Vakkari (Eds.), *Emerging frameworks and methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS 4)* (pp. 137–150). Greenwood Village, CO: Libraries Unlimited.

Committee on Future Roles, Challenges, and Opportunities for the U.S. Geological Survey, National Research Council. (2001). *Future roles and opportunities for the U.S. Geological Survey.* Washington, DC: National Academy Press.

Committee to Review the USGS Coastal and Marine Geology Program, National Research Council. (1999). *Science for decisionmaking: Coastal and marine geology at the U.S. Geological Survey.* Washington, DC: National Academy Press.

Hodge, G. (2000). *Systems of knowledge organization for digital libraries: Beyond traditional authority files.* Washington, DC: Digital Library Federation, Council on Library and Information Resources.

Hutchinson, D.R., Sanders, R., & Faust, T. (2003). *Making USGS information effective in the electronic age* (U.S. Geological Survey Open-File

Report 03-240). Retrieved August 29, 2008, from http://pubs.usgs.gov/of/2003/of03-240/

Janée, G., Frew, J., & Hill, L.L. (2004, May). Issues in georeferenced digital libraries. *D-Lib Magazine, 10*(5). Retrieved August 29, 2008, from http://www.dlib.org/dlib/may04/janee/05janee.html

Lagoze, C. (2001, January). Keeping Dublin Core simple: Cross-domain discovery or resource description? *D-Lib Magazine, 7*(1). Retrieved August 29, 2008, from http://www.dlib.org/dlib/january01/lagoze/01lagoze.html

Lagoze, C., Krafft, D.B., Payette, S., & Jesuroga, S. (2005, November). What is a digital library anymore, anyway? *D-Lib Magazine, 11*(1). Retrieved August 29, 2008, from http://www.dlib.org/dlib/november05/lagoze/11lagoze.html

Lagoze, C., & Van de Sompel, H. (2001). The Open Archives Initiative: Building a low-barrier interoperability framework. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2001)* (pp. 54–62). New York: ACM Press.

Lerner, S., & 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, MA: Woods Hole Oceanographic Institution.

Marincioni, F., Lightsom, F.L., Riall, R.L., Linck, G.A., Aldrich, T.C., & Caruso, M.J. (2004). Integrating digital information for coastal and marine sciences. *Journal of Digital Information Management, 2*(3), 132–141.

Nelson, M.L., & Allen, B.D. (2002, January). Object persistence and availability in digital libraries. *D-Lib Magazine, 8*(1). Retrieved August 29, 2008, from http://www.dlib.org/dlib/january02/nelson/01nelson.html

O'Dea, L., Cummins, V., Wright, D., Dwyer, N., & Ameztoy, I. (2007). *Report on coastal mapping and informatics, Trans-Atlantic Workshop 1: Potentials and limitations of coastal Web atlases*. Cork, Ireland: University College Cork. Retrieved August 29, 2008, from http://workshop1.science.oregonstate.edu/final_rpt

Panel on Distributed Geolibraries, National Research Council. (1999). *Distributed geolibraries: Spatial information resources, summary of a workshop.* Washington, DC: National Academy Press.

U.S. Commission on Ocean Policy. (2004). *An ocean blueprint for the 21st century: Final report.* Washington, DC: U.S. Commission on Ocean Policy.

KEY TERMS

Controlled Vocabulary: A list of preferred terms for indexing information resources, ideally with precise definitions and guidelines for application.

Crosswalk: A semantic mapping between the elements of two metadata standards, facilitating interoperability.

Distributed Geolibrary: In the online environment, a distributed geolibrary allows patrons to search centralized metadata records for information about specific places and then retrieve the original online resources from servers distributed across the Internet.

Faceted Classification: An indexing system that employs several mutually exclusive metadata fields to characterize information resources.

False Drop: An instance of retrieving information that is not relevant to a given search.

Gazetteer: In the traditional sense, a gazetteer is a dictionary of place names. Digital gazetteers, in contrast, link place names to specific bounding boxes or polygonal footprints for the purposes of map display or information retrieval.

Georeferencing: The practice of indexing information resources by geospatial coordinates, place names, or geographic codes.

Granularity: The level of descriptive detail in a metadata record, usually representing a balance between the competing demands of fully characterizing information resources and facilitating the process of search and retrieval.

Interoperability: The ability of different information systems to exchange resources through shared standards.

Knowledge Organization System (KOS): Any formalized scheme for managing information resources, including authority files, subject headings, taxonomies, thesauri, semantic networks, and ontologies.