Distributing and Synchronizing Heterogeneous Metadata for the Management of Geospatial Information Repositories

Elaine L. Westbrooks
Albert R. Mann Library, Cornell University, USA
elw25@cornell.edu

Abstract

Cornell University Geospatial Information Repository, CUGIR, is a Web-based repository providing free access to geospatial data and metadata for New York State. Access is enhanced by transforming and re-purposing geospatial (FGDC) metadata into DC-RDF, Metadata Object Description Schema (MODS) and MARC 21 that is rendered in XML, HTML, SGML, and ASCII text on-the-fly. Although libraries today are increasingly in the position to create and maintain non-MARC metadata, non-MARC resource sharing is largely unchartered territory. This paper will demonstrate how CUGIR exemplifies efficient heterogeneous metadata management while introducing a new model for metadata creation and maintenance typically based in geospatial information system applications. This new model introduces an abridged SODA model as a viable digital library system.
Keywords: *Metadata, CUGIR, FGDC, RDF, SODA, geospatial*

1. Introduction

In 1998, Albert R. Mann Library created the Cornell University Geospatial Information Repository (CUGIR) <http://cugir.mannlib.cornell.edu>, a Web-based repository providing free access to geospatial data and metadata for New York State. Since its beginnings, CUGIR has undergone a series of transformations and upgrades in response to emerging standards and technologies in the field of geospatial information systems (GIS) and digital library research. Continuously adopting new library and GIS standards and developments makes CUGIR increasingly more accessible to users within Cornell University and beyond. The technologies and standards that have been adopted by CUGIR will be described in greater detail in the following paragraphs.

CUGIR possesses a number of characteristics that pose unique challenges for digital library developers. First, most GIS repositories manually distribute data and metadata via CD-ROM, whereas CUGIR freely distributes data and metadata via the WWW, making it a true digital library. Second, CUGIR's invention, support, and subsequent development within an academic research library are rare. In contrast, academic GIS repositories or units are almost always under the jurisdiction of Urban Planning, Architecture, or Geography departments. CUGIR is positioned in a library environment that embraces standards and practices associated with the preservation, retrieval, acquisition, and organization of information. Furthermore, the library community has always been concerned with the archiving and version control of information. It is assumed that consistent application of standards will increase interoperability. It is also assumed that metadata, though costly and difficult, adds value to whatever it describes. While metadata is integral to the administration of CUGIR, the GIS community is most concerned with creating data efficiently, lifting the intense burden of metadata, and distributing data according to user requests. In short, CUGIR reserves a position in two communities, the library and GIS that require the CUGIR team to embrace the standards of both communities. The vast majority of standards from these communities directly impacts metadata and its management. This begs the following questions: if one were to create a perfect and heterogeneous metadata management system for a digital library, namely, CUGIR, what characteristics would it possess? How would it behave? What problems would it solve? The CUGIR team set out to create a system characterized by automatic metadata updating and digital object permanence. The system would behave in a predictable fashion and it would reduce work, costs (automation and less disk space), and increase access. Although the CUGIR metadata model is not a perfect metadata management system, it is efficient. This is largely because it is a hybrid system embracing the standards of the library community while adopting GIS software's most attractive features.

In striving for metadata management perfection, the CUGIR team became keenly aware of the shortcomings, i.e. the lack of version control and preservation, in the way GIS software handles digital objects. The weaknesses of the digital library metadata model (lack of automation) were addressed in two ways. First, the storage of surrogate records for multiple manifestations of the same expression was eliminated. Secondly, the automatic metadata creation tools unique to GIS software applications were used to our advantage. With the weaknesses of both approaches exposed, the team exploited their strengths in order to create more powerful tools. Based upon the team's experiences implementing a metadatabase system, the

author contends that the CUGIR model is a step in the right direction towards improved management of heterogeneous metadata.

Other than Kacmar, Jue, Stage, and Koontz's article "The Automatic Creation and Maintenance of an Organizational Spatial Metadata and Document Digital Library," there is little research and documentation regarding the intellectual and technical organization and management of spatial metadata [1]. Many governmental and private agencies have published grey literature regarding the principles of good metadata management, but principles do not constitute methods or even best practices [2; 3].

The purpose of this paper is to introduce a new metadata management model. This model specifically attempts to address the following problems:

a) managing multiple metadata schemas in multiple manifestations and expressions in digital libraries;
b) the lack or absence of centrality, persistence/permanence of (geospatial) digital objects in digital libraries;
c) the creation and maintenance of metadata that is almost always difficult, costly, and time consuming; and,
d) the lack of metadata synchronization tools in traditional and digital libraries.

It was the goal of the CUGIR team to take the best of both worlds (digital libraries and GIS applications) and merge them to make a powerful system from which both communities could benefit and of which they could be proud. Although this model was chiefly designed for geospatial data and metadata, it is applicable to other types of digital libraries. Before this model is presented, it is best to view it in context, beginning with the history of CUGIR.

2. CUGIR History

CUGIR is a clearinghouse and repository that provides unrestricted access to geospatial data and metadata, with special emphasis on those natural features relevant to agriculture, ecology, natural resources, and human-environment interactions in New York State. Staff at Albert R. Mann Library at Cornell University began looking at ways to disseminate geospatial data from Mann's collections via the WWW in 1995, and in 1998 established a Web-based clearinghouse for New York State geospatial data and metadata. Building a clearinghouse entailed creating partnerships with local, state and federal agencies, understanding how to interpret and apply the Federal Geographic Data Committee (FGDC) Content Standard for Geospatial Metadata (CSDGM), and designing a search and retrieval interface, as well as a flexible and scalable data storage system [4].

The development of CUGIR has been accomplished through a team-based model of work and cooperation. The CUGIR team were identified and selected from departments within Mann Library: Public Services, Technical Services, Collection Development, and Information Technology. This team provides for the management, preservation, organization, and storage needs of datasets which are distributed in CUGIR, but which are owned by various departments in New York State governmental agencies as well as Cornell-affiliated departments, agencies, and researchers [4]. The CUGIR team consists of five regular members, each coordinating work within their areas of specialty. Other library staff participated on an as-needed basis. Primary responsibilities for the overall coordination of clearinghouse development are carried out by the GIS librarian.

CUGIR is one of 250 international nodes within the Geospatial Data Clearinghouse that contain searchable metadata records describing geospatial datasets. All nodes are located on data servers using the Z39.50 information retrieval protocol. As a result, nodes can be linked to a single search interface where the metadata contents of all nodes, or any subset in combination, can be searched simultaneously. CUGIR, like most clearinghouse nodes, has its own Web site with customized browsing and searching interfaces [4; 5]. Statistics indicate that CUGIR's utility and popularity continues to grow. Since 1998, CUGIR data requests have increased by at least forty percent each year. In fact, it is projected that CUGIR will record over 100,000 requests in 2003, the most for any single year since the repository was established in 1998 [6].

2.1. CUGIR Data and Datasets

Currently, CUGIR freely distributes online over 7,000 datasets produced by ten data producers or partners (Cornell Department of Natural Resources, NY Department of Energy Conservation, Soils Information System Laboratory, NY Department of Agriculture and Markets, Adirondack Park Agency, NY Tompkins County GIS, USGS, US Census, National Atlas, USDA). CUGIR data come in seven unique proprietary and non-proprietary formats (ArcExport, shapefile, CAD, geoTIFF, PDF, ArcInfo Grid, DEM) [7]. In many cases, one dataset is produced in multiple formats. For example the dataset: "Minor Civil Divisions, Albany County" is available in ArcExport as well as shapefile format. Each format has unique characteristics that make it more or less desirable for certain uses and purposes. CUGIR data are actively maintained according to the needs of the data producers.

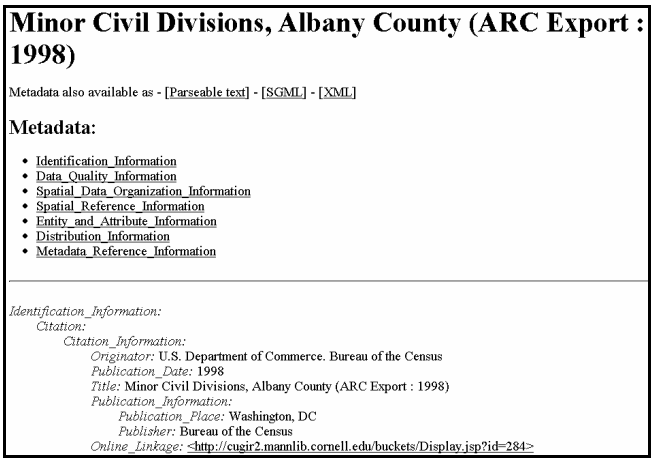
Unlike most digital library files that require little more than Internet connectivity and Web browser software, geospatial data require technical expertise in the use of sophisticated and powerful GIS software applications. In addition, users must also understand cartographic and geographic concepts related to GIS.

2.2. CUGIR FGDC Metadata

In 1994, the FGDC established the CSDGM for describing the content and function of geospatial data. According to the FGDC, "The standard was developed… to determine the availability of a set of geospatial data, to determine the fitness of a set of geospatial data for an intended use, to determine the means of accessing the set of geospatial data …. As such, the standard established the names of data elements and compound elements to be used for these purposes, the definition of these data elements and compound elements to be used for these purposes" [8]. All data producers should provide up-to-date and accurate information about what data are available and their characteristics. The collection, management, and distribution of *good* metadata can help achieve this goal [3]. A high percent of CUGIR metadata is produced by the data producer and all of it is summarily reviewed and enhanced by the metadata librarian.

There are 334 different elements in FGDC's CSDGM, 119 of which exist only to contain other elements [9]. These elements are organized within seven main sections and three supporting sections that describe different aspects of data that potential users might need to know: Identification Information; Data Quality Information; Spatial Data Organization Information; Spatial Reference Information; Entity and Attribute Information; Distribution Information and Metadata Reference Information. Of these areas only Identification Information (basic information about the file such as originator, abstract, and purpose) and Metadata Reference Information (information about the production of the metadata) are defined as being mandatory for all records. All other areas of the standard are mandatory if applicable. Within each section are sub-fields that can be defined as mandatory, mandatory if applicable, or optional. This flexibility allows metadata creators to determine the level of detail that they can provide or support based on perceived user needs. It also guarantees that at least basic metadata will be recorded about each dataset. For more extensive information about FGDC metadata creation, see Hart and Phillips' Metadata Primer [10].

CSDGM is extremely detailed, hierarchical, and complex, which explains why many organizations fear it. In the case of CUGIR, that active management and quality control of metadata is the responsibility of the team's metadata librarian. The metadata librarian enhances and edits the metadata to make it FGDC-compliant. Figure 1 is an example of an FGDC CUGIR record entitled, "Minor Civil Divisions, Albany County." The "Online_Linkage" element, links users to the Dublin Core (DC) record where the data can be downloaded. This special kind of link constitutes a central digital object known as a bucket. This concept will be discussed in detail under section 2.4--Buckets: Smart Object Dumb Archive (SODA).



Minor Civil Divisions, Albany County (ARC Export : 1998)

Metadata also available as - [Parseable text] - [SGML] - [XML]

Metadata:

- Identification_Information
- Data_Quality_Information
- Spatial_Data_Organization_Information
- Spatial_Reference_Information
- Entity_and_Attribute_Information
- Distribution_Information
- Metadata_Reference_Information

Identification_Information:
Citation:
Citation_Information:
Originator: U.S. Department of Commerce. Bureau of the Census
Publication_Date: 1998
Title: Minor Civil Divisions, Albany County (ARC Export : 1998)
Publication_Information:
Publication_Place: Washington, DC
Publisher: Bureau of the Census
Online_Linkage: <http://cugir2.mannlib.cornell.edu/buckets/Display.jsp?id=284>

Figure 1. Geospatial/FGDC metadata record in CUGIR. From this record, one may download the dataset from the Online Link

Of the 7117 datasets in CUGIR, ninety-nine percent (7111) are accompanied by FGDC-compliant metadata. All metadata is reviewed and enhanced by the metadata librarian before the data and metadata are added to CUGIR. CUGIR metadata are created and stored as ASCII text, HTML, SGML, and XML. Online users may view any metadata record in any format of their choice.

It is worth emphasizing that geospatial metadata and data come with a host of issues that distinguish them from most digital objects. Spatial data files are complex objects and it is difficult to construct locator and description records for them without the use of specialized searching tools. For example, traditional Boolean word operations are not optimal for determining whether a spatial object is relevant to a particular task. Three dimensional search engines are necessary to search data having three primary elements: attributes, time, and user tasks. Libraries are accustomed to adhering to standards, yet, the vast majority of GIS repositories are not managed in libraries. As a consequence, there are few digital libraries that take geospatial metadata into consideration. This lack of development (the exception to this rule is the Alexandria Digital Library at the University of California at Santa Barbara) and research in geospatial digital libraries has made geospatial research and metadata development forever challenging and frequently groundbreaking. Moreover, it is this reality that has forced the CUGIR team to strive for a framework that not only enhances access and shares heterogeneous metadata, but also fosters digital object permanence and centrality in a way that makes the metadata management more efficient, cost-effective, and interoperable. In advancing the concept of digital libraries, the CUGIR team affirms Jane Greenberg's statement, "the success of digital libraries, interoperability, and evolution of the semantic web all rely on efficient metadata generation" [11]

2.3. CUGIR Metadata Management

In a broad sense, and in the case of CUGIR, metadata management, by definition, implies the implementation of a metadata policy [3] (i.e. principles that form the guiding framework within which metadata exists) and adherence to metadata standards. Furthermore, metadata management is the process of acquiring and maintaining a controlled set of metadata in order to describe, discover, retrieve, and access the data to which it refers. The more complex, relational, and heterogeneous CUGIR metadata became, the more it became necessary to have a management system that could deal with the known problems: access and redundancy.

The CUGIR team identified one major area essential to CUGIR's success—access. Cornell University's core constituency of faculty, students, and staff-- were clearly not utilizing CUGIR's geospatial resources. Metadata records were not fully accessible, residing inside the CUGIR Website and the NSDI which both occupy the "Deep Web [12]." For the team, the question became, How do we make geospatial information resources more accessible to users who might not otherwise encounter them? Because complex metadata schemas like MARC 21 and FGDC are not the 'languages' of the WWW, it became clear that more accessible metadata standards must be used to increase CUGIR's web presence in spite of the deep Web. At the same time, MARC, which is not a language for the WWW, remains the most prominent and reliable metadata schema for libraries today, and potentially tomorrow. Consequently, all FGDC records were converted to MARC for the online catalog (OPAC), as well as other metadata schemas for sharing and distribution throughout a number of metadata management systems.

Another identified problem was the prevalence of redundant metadata records that differ only in format. The storage of metadata in HTML, XML, SGML, and ASCII text was difficult to manage when changes were necessary. Similarly, the repetition of metadata elements or fields in those metadata also demonstrated inefficient use of storage space. In order to address these problems, the CUGIR team set out to introduce a more accessible and efficient management system, centered on one metadata work in particular, the canonical record.

2.3.1. Canonical CUGIR Metadata

In order to minimize the amount of data lost as a result of crosswalking among multiple schemas, the metadata schema conversion process began with the core, or canonical FGDC record which is assembled on-the-fly. The FGDC record is considered the "native" and most complete source of information, in one of the most flexible exchange formats, XML. With no existing tools to convert FGDC XML to MARC XML, this was quite a challenge. Elizabeth Mangan, of The Library of Congress (LC), created a FGDC to MARC 21 crosswalk that was useful, but a new and customized FGDC XML to MARC XML crosswalk had to be created to suit our purposes [13; 14]. The MARC XML is also derived from the canonical form and produced on-the-fly.

What makes the use of the canonical record even more important is the upcoming introduction of ISO geospatial metadata. ISO metadata when implemented will harmonize the FGDC Metadata Standard (FGDC-STD-001-1998) with ISO's Geographic Information/Geomatics Technical Committee (TC) 211 Metadata Standard 19115. The standard will be a multilingual XML schema designed to be extensible (profile and extension friendly), multi-layered (supporting relational hierarchy of metadata), and modeled in Unified Modeling Language (UML). In addition, it will be integrated with other ISO standards such as DC (ISO 15836) and Codes for the Representation Languages Names (ISO639-2). This harmonization process is a powerful step in the right direction because it not only addresses many known deficiencies in FGDC CSDGM, but also enables interoperability while providing additional support for the functions of metadata. Embracing XML encoded FGDC is the CUGIR team's way of dealing with the upcoming changes. Given the metadata tools and practices we have in place, we expect a predictable and effortless transition from FGDC to ISO. Thus CUGIR will be poised to make the transition, instead of waiting for proprietary metadata tools to emerge.

In order to minimize the storage of redundant information, the canonical record is stored in a database and produced on-the-fly. For example, each data partner has standard contact information (e.g. address, telephone number) that is recorded in every metadata record. Instead of repeating such information in each and every metadata record, it is stored once and produced on-the-fly. Figure 2 below illustrates the CUGIR metadata conversion process.

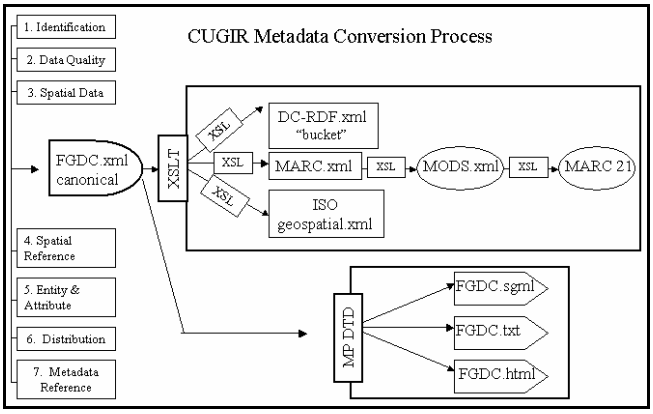


Figure 2 CUGIR Metadata Conversion Process beginning from the left to the right

2.3.2. DC-RDF for OAI and the Semantic Web

The online repository (other than CUGIR itself) chosen to increase access to CUGIR was the Open Archives Initiative (OAI) Community. OAI develops and promotes interoperability standards that aim to facilitate the efficient

dissemination of content [15]. In addition, The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) provides an application-independent interoperability framework based on metadata harvesting [16]. The recommended, but not required, metadata schema for the OAI, is Dublin Core (DC). The CUGIR team chose to use DC and the Resource Description Framework (collectively known as DC-RDF) for a number of reasons, the first being the convenient use of OCLC's Connexion to export OAI-ready DC-RDF with little effort. The methods by which the DC-RDF records were produced will be further discussed below in section 2.4. As the metadata project progressed, we favored a less OCLC-centric approach to metadata creation. Moreover, we discovered that DC-RDF metadata records (in XML) could be easily created with XML stylesheets (XSL) coupled with extensible stylesheet language transformations (XSLT). XSL defines how data are presented while XSLTs are designed for use as part of XSL. DC-RDF is naturally encoded in XML, which is an exchange format through which that data providers harvest and share metadata. Another attractive feature of the OAI-PMH is it's use of HTTP over the complex information retrieval protocol Z39.50. Although Z39.50 has served the library community well, the simplicity of having servers provide CUGIR metadata in bulk for harvesting services by way of HTTP is a viable alternative to National Spatial Data Infrastructure (NSDI) currently in place for GIS repositories across the globe.

Chandler and Foley's 2000 study documents the problems inherent in gaining access to spatial data via the NSDI [17]. NSDI nodes are inconsistently available, searches are inaccurate, and the wait time is excessive. In his article "Metadata Harvesting and the Open Archives Initiative" Clifford Lynch, the executive director of the Coalition for Networked Information, also documents the strengths and weaknesses of Z39.50, in the context of OAI, [18]. The use of RDF can be easily justified when one considers the integral role it performs in the Semantic Web [18]. According to Tim Berners-Lee, "The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is based on the RDF, which integrates a variety of applications using XML for syntax…"[19]. Eric Miller and Ralph Swick of the World Wide Web Consortium report, "for the Web to reach its full potential, it must grow and incorporate a semantic Web vision, providing a universally accessible platform that allows data to be shared and processed by people and machines" [20]. Thus by embracing RDF the CUGIR team aims to situate CUGIR metadata schemas in a position to flourish within the semantic Web. Finally, up and coming information management systems such as D-Space, Open Archival Information Systems (OAIS), EnCompass, and ExLibris use some form of DC encoded in XML as the lingua franca. We assume that CUGIR will be distributed in such systems in the near future.

2.3.3. Metadata Management: MARC

The contribution of MARC 21 records to OCLC makes CUGIR data internationally accessible to WorldCat users. Additionally, other libraries on the OCLC network get the opportunity to utilize full level MARC records [21]. The integration of CUGIR data into the OPAC made it possible for library users to discover geospatial resources as they typically discover journals, books, and online databases. In sum, the transformation from FGDC to MARC 21 enabled the CUGIR team to do the following:

a) Gain bibliographic control over CUGIR records;
b) Enhance access to geospatial records via the OPAC; and
c) Share MARC 21 records with libraries worldwide via WorldCat.

The coexistence of geospatial metadata with traditional resources in the OPAC is essential to making geospatial datasets known and accessible beyond the narrow world of GIS [22]. MARC 21 is based on the XML encoded FGDC records and transformed on-the-fly using XSLT. Concurrently, the MARC 21 records are added to the OPAC in a batch process.

While we are already creating multiple metadata schemas on-the-fly it seems only natural that we include some of the latest developments in metadata. Though not thoroughly tested, they display great potential and innovation. The Metadata Object Description Schema, MODS is a subset of MARC 21 and one of the latest developments worthy of investigation. According to its official Website, "As an XML schema, MODS, is intended to be able to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records. It includes a subset of MARC fields and uses language-based tags rather than numeric ones, in some cases regrouping elements from the MARC 21 bibliographic format. MODS is expressed using the XML schema language of the World Wide Web Consortium" [23]. Rebecca Guenther, LC Senior Networking and Standards Specialist, adds that "MODS should complement other metadata formats and should provide an alternative between a very simple metadata format with a minimum of fields and no or little substructure (i.e. DC) and a very detailed format with many data elements having various structural complexities such as MARC 21" [24].

The adoption of MODS into the metadata framework required the metadata librarian to build a FGDC to MODS crosswalk, stylesheet, and transformation, since none existed [13]. There are a few institutions other than LC and the California Digital Library that are currently producing MODS records. It is safe to assume that MODS will become one of the sanctioned metadata schemas of the OAI MHP in the near future. MODS is an attractive XML descriptive standard, particularly in the way it provides flexibility and can be combined with other XML-based standards including the Metadata Encoding Transmission

Schema (METS). The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium [25]. To quote Guenther and McCallum, "An additional use of MODS is as an extension schema for descriptive metadata for METS objects…" [26] Since any descriptive metadata that is part of CUGIR can be part of METS objects, we anticipate that our next step will be to investigate how well METS can handle geospatial information. Presumably, if we are satisfied, the presence of MODS will help CUGIR transition gracefully into METS.

2.4. Buckets: Smart Object Dumb Archive (SODA)

Creating multiple schemas on-the-fly from XML encoded FGDC was one of the easiest steps in the metadata project. However, adopting a simple method to create, maintain, and centralize the persistent/permanent hyperlinks in metadata proved to be a formidable challenge. John Kunze, researcher at the University of California at San Francisco Library Center for Knowledge Management, articulates the problem, stating, "Permanence of electronic information, namely, the extent to which structured digital data remains predictably available through known channels, is a central concern for most organizations whose mission includes an archival function" [27]. The difficulty in carrying out the aim of permanence, and indeed, centrality in the CUGIR digital library system is in identifying existing solutions that are more simple, flexible, and dynamic than your everyday Universal Resource Locators (URL) or Persistent Uniform Resource Locators (PURL) heavily used in digital libraries and OPACs today. Given the inadequacies of URLs and PURLs, it became apparent to the CUGIR team that we needed an identifier resolver that de-couples the identity of the object from the location of the object, while providing more functionality. The solution was a complex resolver known as a "bucket". The term "bucket" is borrowed from Michael Nelson's research on digital library architecture [28]. To quote Nelson, professor of Computer Science at Old Dominion University, "Buckets are a part of the larger "Smart Object Dumb Archive" [Digital Library] Model. SODA is a reaction to the vertically integrated (and non-interoperable) DLs that tended to grow…Separating the functionality of the archive from that of the DL allows for greater interoperability…"[29]. In another article, he states further, "Buckets are object-oriented container constructs in which logically grouped items can be collected, stored, and transported as a single unit" [30]. For the digital library, the bucket became the glue that held the metadata framework together, by enabling identifier persistence across the heterogeneous metadata surrogates of FGDC, DC-RDF, and MARC records distributed in CUGIR, the OPAC, OCLC (WorldCat), as well OAI. In short, Nelson's bucket solution was a way to group everything (i.e., metadata) in a common place, building a small container around it. For our purposes, only part of Nelson's bucket architecture was sufficient to meet the needs of the metadata framework.. Therefore, we arrived at a system that borrowed the simplicity of a PURL for resolving identifiers, but added the capacity of linking related objects together into a coherent framework. We refer to our system as the CUGIR Simple SODA Model, illustrated in Figure 3. Let us take a look at the composition of the bucket.

The bucket is composed of three pieces of information, the data theme, the mapsheet, and the dataset format.

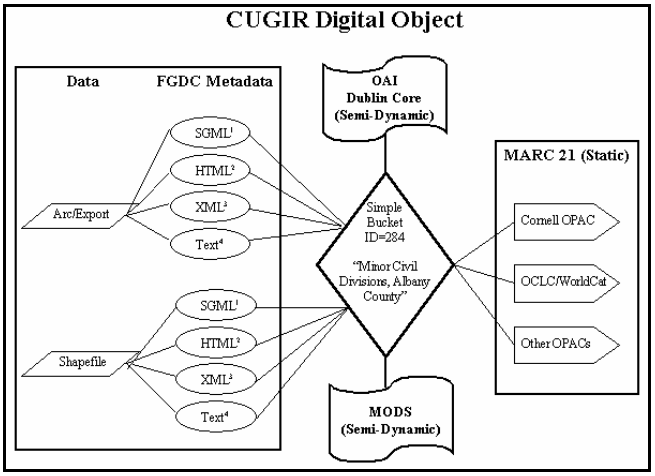
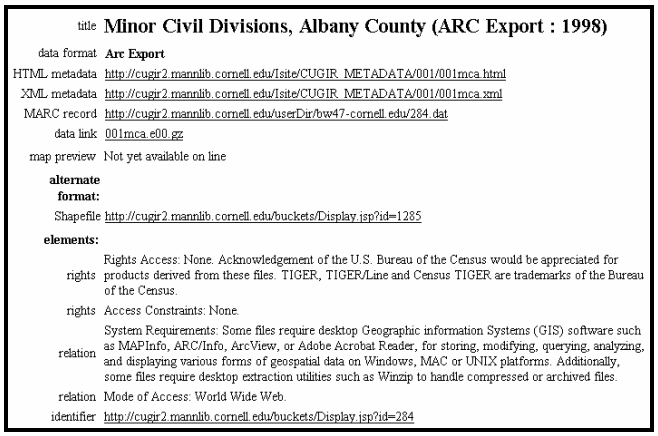


Figure 3. CUGIR Bucket Digital Object Model

The CUGIR digital object has three components. Working from left to right in Figure 3, the "CUGIR" box represents the original FGDC metadata files, in all four formats (SGML, HTML, XML, and ASCII text) plus the dataset described by the metadata. The bucket in the middle binds the digital objects together. It contains the location of the different CUGIR metadata files. It does not need to contain the location of the data file, since the data file location is maintained within the CUGIR metadata file. The bucket location on the CUGIR network is by design stable and persistent, like a PURL, thus creating the possibility of linking to CUGIR metadata from MARC surrogates placed in Cornell's OPAC, and OCLC's WorldCat. These records also contain links to buckets [21]. The design and implementation of the SODA system was financed by the Cornell University Libraries [31].

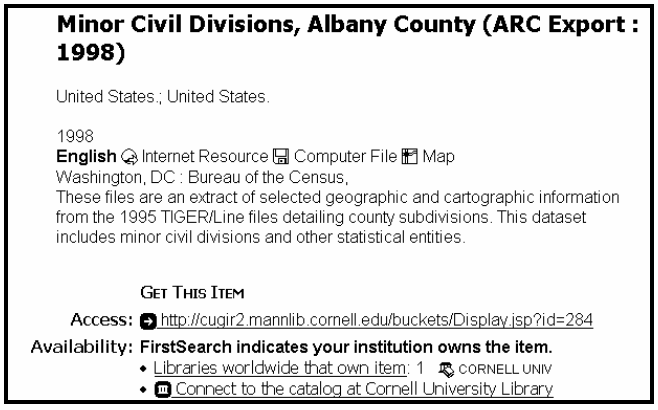
Figure 4 is an example of a bucket that is rendering DC-RDF for "Minor Civil Divisions, Albany County". From the bucket the user has access to the full CUGIR geospatial metadata record labeled "HTML Metadata" or "MARC record" respectively. For all intents and purposes, the DC-RDF is the bucket.



title	**Minor Civil Divisions, Albany County (ARC Export : 1998)**
data format	**Arc Export**
HTML metadata	http://cugir2.mannlib.cornell.edu/Isite/CUGIR_METADATA/001/001mca.html
XML metadata	http://cugir2.mannlib.cornell.edu/Isite/CUGIR_METADATA/001/001mca.xml
MARC record	http://cugir2.mannlib.cornell.edu/userDir/bw47-cornell.edu/284.dat
data link	001mca.e00.gz
map preview	Not yet available on line
alternate format:	
Shapefile	http://cugir2.mannlib.cornell.edu/buckets/Display.jsp?id=1285
elements:	
rights	Rights Access: None. Acknowledgement of the U.S. Bureau of the Census would be appreciated for products derived from these files. TIGER, TIGER/Line and Census TIGER are trademarks of the Bureau of the Census.
rights	Access Constraints: None.
relation	System Requirements: Some files require desktop Geographic information Systems (GIS) software such as MAPInfo, ARC/Info, ArcView, or Adobe Acrobat Reader, for storing, modifying, querying, analyzing, and displaying various forms of geospatial data on Windows, MAC or UNIX platforms. Additionally, some files require desktop extraction utilities such as Winzip to handle compressed or archived files.
relation	Mode of Access: World Wide Web.
identifier	http://cugir2.mannlib.cornell.edu/buckets/Display.jsp?id=284

Figure 4. Online Bucket #284, which embodies the DC-RDF describing "Minor Civil Divisions, Albany County"

Figure 4 demonstrates how WorldCat users searching for "Minor Civil Divisions, Albany County" will arrive at the same bucket when they click on the hyperlink in the record which eventually leads patrons to the CUGIR canonical record in FGDC as seen in Figure 1 on page 3 of this paper.



Minor Civil Divisions, Albany County (ARC Export : 1998)

United States.; United States.

1998
English Internet Resource Computer File Map
Washington, DC : Bureau of the Census,
These files are an extract of selected geographic and cartographic information from the 1995 TIGER/Line files detailing county subdivisions. This dataset includes minor civil divisions and other statistical entities.

GET THIS ITEM

Access: http://cugir2.mannlib.cornell.edu/buckets/Display.jsp?id=284

Availability: **FirstSearch indicates your institution owns the item.**
- Libraries worldwide that own item: 1 CORNELL UNIV
- Connect to the catalog at Cornell University Library

Figure 5 MARC record in WorldCat. Notice how the "Access" (MARC 856) field leads back to bucket #284.

What is most valuable about this framework is the fact that information in thousands of metadata records can be easily changed. For example, if CUGIR were moved to another server, all of the URLs would no longer work. Presumably, the buckets, which constitute the 856 field of the MARC record as well as the "Online_Linkage" field of the FGDC record would have to be changed in the metadata residing in the OPAC, OCLC, and OAI as well as the canonical FGDC record in CUGIR. In CUGIR's SODA system, however, only the bucket requires changing, as opposed to all of the records in the OPAC, OCLC, and OAI. This model is efficient for the CUGIR team and eliminates the need to update metadata records every time CUGIR metadata are moved to new servers.

In the ultimate metadata management system, the *content* of the canonical FGDC record and its derivatives (DC-RDF, MARC 21, MODS) would be automatically updated when the dataset that is being described is altered. Presently GIS software applications use an automatic updator and creator known as the metadata synchronizer. We now turn to the metadata editor that creates and synchronizes metadata.

2.5. Metadata Editing and Automatic Metadata Creation/Synchronization

CUGIR currently uses a number of Environmental Systems Research Institute (ESRI) software, commonly used in geospatial information analysis, to manage and store CUGIR data and metadata. This suite of services includes ArcGIS, an Internet Mapping Service (ArcIMS), and a Spatial Data Engine (ArcSDE). ArcGIS contains a data management tool known as ArcCatalog, which is a data exploration and management application. ArcCatalog is used to preview metadata as well as a dataset's geographic and tabular data. The most attractive features of ArcCatalog are the metadata editor and creator.

ArcCatalog includes a FGDC-compliant metadata editor that creates metadata records using any or all of the elements defined in the CSDGM. Metadata created with ArcCatalog are stored as XML and indexed within the CUGIR geodatabase. ArcCatalog comes with stylesheets that produce XML encoded FGDC and ISO geospatial metadata. The CUGIR team added stylesheets that produce FGDC metadata encoded in SGML, HTML, and ASCII text as well as FGDC represented in the schemas of our choice, DC-RDF, MARC XML, and MARC 21.

ArcCatalog automatically creates metadata for datasets stored in the geodatabase if none exists. Some of the automatically generated metadata describe the dataset's current properties, i.e coordinate system, entity, and attribute information. Every time the metadata librarian views the metadata, ArcCatalog automatically updates or synchronizes dataset properties with its most current values. Of course, the synchronization ensures that the metadata is perpetually up to date according to the changes in the dataset. Synchronization is accomplished as a result of the values held by the Sync attribute. When the Catalog initially records a dataset's properties in the metadata, the Sync attribute for the associated element is "TRUE". When ArcCatalog updates metadata, if it does not find the Sync attribute or its value is not "TRUE", it will not overwrite the element's value. Automatic synchronization is an invaluable feature, but it brings forth a host of problems associated with archiving and bibliographic control. That is to say, making distinctions between and among metadata versions, editions, and updates is crucial for any type of digital library with archiving responsibilities such as CUGIR. The inability of the synchronizer to differentiate a version of a metadata record from an edition, or update brought forth a new set of challenges.

2.6. Outcomes of CUGIR Metadata Framework

The CUGIR metadata framework proved successful in reaching its primary goals: increasing access and implementing an efficient metadata management system. But, what impact did all of this work have on CUGIR's users? In other words, did more Cornell constituents discover CUGIR as a result of the metadata framework? The answer to this question is "yes".

When the framework was implemented, referrer data was captured so as to indicate the Webpage that a user visited in order to access the bucket. The IP addresses of the hosts were also collected. To preserve the privacy of users, the IP addresses have been encrypted and the subnets were dropped from the statistics database. As a result, the domain name rather than the unique address of the computer have been stored. These data confirm when users encountered a bucket from OAI, the OPAC, or FirstSearch. We established a tracking method that observes use patterns and indicates the manner and frequency with which patrons access buckets. Since the metadata framework has been in place approximately 12,000 buckets have been accessed from a variety of locations. Unfortunately, we do not have enough data about the OAI user's harvesting of CUGIR DC-RDF records. The results indicate that less than five percent of our users discover CUGIR metadata via the OPAC. Less than one percent of our users discover CUGIR metadata via FirstSearch. Almost ninety-five percent of our users discover CUGIR metadata from CUGIR's homepage.

If only five percent of our users discovered CUGIR as a result of this metadata framework, was it worthwhile? Although the statistics do not indicate "success", in regard to access, the work and process of formulating the metadata sharing framework forced us to document all metadata processes, streamline workflows, and create more metadata with less effort. In terms of data management, the metadata framework reduced the number of metadata files that had to be managed and stored. CUGIR no longer stores each metadata schema in multiple formats. In the past, we stored nine metadata files per dataset. Now we only store one. GIS is a growing field that is increasingly being used across disciplines in the academy and in ninety-five percent of all government planning decisions [32].

4. Conclusion

We are confident that our work to make CUGIR more accessible will pay off in the long run. Furthermore, the proliferation of Web Mapping services will expose GIS to even more users who might not otherwise know about it. Increasingly diverse and sophisticated Web sites today allow instant creation of customized maps. Such interactive mapping Web sites exemplify the most dynamic aspects of GIS. As the spatial Web grows, there are more online spatial resources available. And while these resources are getting simpler to use, there is increasing potential for extended capability and complexity in Web mapping applications. As Web sites become richer in processing resources, users will need to own less GIS software and their sessions on the Web will become more interactive. With Web mapping, users can view and access data online without having expensive software (e.g. ArcGIS), a complete understanding of GIS technology, or expertise in the cartographic and geographic concepts related to GIS. Many repositories are beginning to offer interactive mapping sites where one can create maps based on huge census, EPA, or the USGS databases of information. Finally, the worth of the CUGIR metadata framework is evident from the growing importance of standards in the GIS community. Consortia such as the Open GIS Consortium are aimed at growing interoperability for technologies involving spatial information and location so that everyone benefits from geographic information and services made available across any network, application, or platform [33].

Thus, the data analysis of the use of the CUGIR metadata management system yielding some interesting insights:

a) In spite of the vast efforts to make CUGIR data accessible across metadata schemas and information systems, users who know about CUGIR overwhelmingly prefer to acquire data from the FGDC metadata records on the CUGIR Homepage. This will always be the case no matter how much metadata sharing persists;

b) The OPAC provides minimal means for access for a set of users who might not otherwise discover geospatial data; and,

c) If the SODA system and the metadata framework did not make metadata records so easy to create and maintain, then we would not make the effort to contribute data to OCLC's FirstSearch. The addition of MARC 21 records in OCLC has not significantly increased access to CUGIR. On the other hand, other libraries in the OCLC network have access to full level MARC records and may find them useful.

The fundamental value of the library is the organization of information as the foundation through which information resources can be utilized. Centuries of library research support this claim. The same principles are not being applied to digital libraries. The CUGIR team embraces metadata as the first-order prerequisite to establishing a complete spatial repository or clearinghouse as well as the Semantic Web. Further more, it should be clear that library standards and theory as well as GIS standards and software must be applied in concert, in order to produce open, interoperable, efficient, and robust digital libraries.

Acronyms

ArcGIS- ESRI's Arc Geospatial Information System software
ArcIMS- ESRI's Arc Internet Mapping Service software
ArcSDE- ESRI's Arc Spatial Data Engine software
ASCII- American Standard Code for Information Interchange
CAD- Computer Aided Design
CSDGM- Content Standard for Digital Geospatial Metadata
CUGIR- Cornell University Geospatial Information Repository
DC- Dublin Core
DC-RDF- Dublin Core Resource Description Framework
DEM- Digital Elevation Models
EPA- Environmental Protection Agency
ESRI- Environmental Systems Research Institute
FGDC- Federal Geographic Data Committee
GIS- Geographic Information System
HTML- Hypertext Markup Language
HTTP- Hypertext Transfer Protocol
ISO- International Standards Organization
ISO 15836- International Standards Organization Dublin Core Metadata Element Set Number
ISO 19115:2003- International Standards Organization Metadata Schema for Geospatial Metadata Number
ISO TC211- International Standards Organization Technical Committee for Geographic Information/Geomatics
MARC21- MAchine Readable Cataloging
METS- Metadata Encoding & Transmission Schema
MODS- Metadata Object Description Schema
NSDI- National Spatial Data Infrastructure
NY- New York
OAI- Open Archives Initiative
OAI-PMH- Open Archives Initiative Protocol for Metadata Harvesting
OAIS- Open Archival Information System
OCLC- Online computing Library Center
OPAC- Online Public Access Catalog
PDF- Portable Document Format
SGML- Standard Generalized Markup Language
SODA- Smart Object Dumb Archive
URL- Uniform Resource Locator
USDA- United States Department of Agriculture
USGS- United States Geological Survey
XSL- eXtensible Stylesheet Language
XSLT- eXtensible Stylesheet Language Transformation
XML- eXtensible Markup Language
Z39.50- Application Service Definition and Protocol Specification for Information Retrieval

Acknowledgements

I would like to express my gratitude to Adam Chandler whose groundwork and collaboration led to the CUGIR metadata framework as we know it. A special thanks to Jonathan Corson-Rikert, Gretchen Higginbottom, Jaime Martindale, Suzette Spencer, and Marijo Wilson, for their valuable feedback.

References

[1] Kacmar, C., Jue, D., Stage, D. & Koontz, C. (1995). Automatic Creation and Maintenance of an Organizational Spatial Metadata and Document Digital Library. Retrieved May 3 2003, from The Second Annual Conference on the Theory and Practice of Digital Libraries, June 11-13, 1995 - Austin, Texas, USA Proceedings Website: http://www.csdl.tamu.edu/DL95/papers/kacmar/kacmar.html.

[2] Westcott, B. (2002). Spatial Metadata for Management -- Increasing the Value of your Data Investment. Retrieved December 1 2002, from PDF Document: http://tsc.wes.army.mil/tsc2/symposium/2002/276.pdf.

[3] IGGI. (2002). Principles of Good Metadata Management. Retrieved July 17 2003, from Intra-governmental Group on Geographic Information Working Group on Metadata Implementation Guide: http://www.iggi.gov.uk/achievements_deliverables/pdf/Guide.pdf.

[4] Herold, P., Gale, T. D. & Turner, T. P. (1999). Optimizing Web Access to Geospatial Data: The Cornell University Geospatial Information Repository (CUGIR). Retrieved January 24 2003, from Issues in Science and Technology Librarianship Online Journal: http://www.library.ucsb.edu/istl/99-winter/article2.html.

[5] Herold, P., Turner, T. P. & Gale, T. D. (1999). Final Project Report: Cornell University Geospatial Information Repository (CUGIR).

[6] CUGIR. (2003). CUGIR Statistics Database. Retrieved May 1 2003, from CUGIR Website: http://rikert.mannlib.cornell.edu/cugir/jsp/downloads.jsp.

[7] CUGIR. (2003). CUGIR Help Files. Retrieved April 5 2003, from Cornell University Geospatial Information Repository Web site: http://cugir.mannlib.cornell.edu/help/help.html.

[8] FGDC. (n.d.). Federal Geographic Data Committee: Metadata (CSDGM). Retrieved March 15 2001, from FGDC Website: http://www.fgdc.gov/metadata/metadata.html.

[9] Schweitzer, P. (n.d.). Frequently-Asked Questions on FGDC Metadata. Retrieved February 2 2003, from FGDC Website: http://geology.usgs.gov/tools/metadata/tools/doc/faq.html.

[10] Hart, D. & Phillips, H. (2001). Metadata Primer -- "How To" Guide on Metadata Implementation. Retrieved August 10 2001, from http://www.lic.wisc.edu/metadata/metaprim.htm.

[11] Greenberg, J. (2002/2003). Metadata Generation: Processes, People and Tools. Bulletin of the American Society for Information Science and Technology, 29(2): 16-19.

[12] Bergman, M. K. (2001). The Deep Web: Surfacing Hidden Value. Retrieved April 5 2002, from University of Michigan Press Online Journal 7(1): http://www.press.umich.edu/jep/07-01/bergman.html.

[13] Westbrooks, E. L. (2003). FGDC to MODS Crosswalk. Retrieved April 30 2003, from Website: http://metadata-wg.mannlib.cornell.edu/elaine/fgdc/.

[14] Mangan, E. (1997). Crosswalk: FGDC Content Standards for Digital Geospatial Metadata to USMARC. Retrieved December 12 2000, from Alexandria Digital Library Website: http://alexandria.sdc.ucsb.edu/public-documents/metadata/fgdc2marc.html.

[15] Lagoze, C., Van de Sompel, H., Nelson, M. L. & Warner, S. (2002). Open Archives Initiative Frequently Asked Questions (FAQ). Retrieved March 1 2003, from Open Archives Initiative Website: http://www.openarchives.org/documents/FAQ.html.

[16] Lagoze, C., Van de Sompel, H., Nelson, M. L. & Warner, S. (2001). The Open Archives Initiative Protocol for Metadata Harvesting. Retrieved March 1 2003, from Open Archives Initiative Website: http://www.openarchives.org/OAI/openarchivesprotocol.html.

[17] Chandler, A. & Foley, D. (2000). Mapping and Converting Essential Federal Geographic Data Committee (FGDC) Metadata into MARC21 and Dublin Core: Towards an Alternative to the FGDC Clearinghouse. Retrieved December 7 2000, from D-Lib Magazine Online Journal: http://www.dlib.org/dlib/january00/chandler/01chandler.html.

[18] Lynch, C. (2001). Metadata Harvesting and the Open Archives Initiative. Association of Research Library Bimonthly Report, 217: 1-9.

[19] Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web. Retrieved February 12 2003, from Scientific American Online Magazine: http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2.

[20] Miller, E. & Swick, R. (2003). An Overview of W3C Semantic Web Activity. Bulletin of the American Society for Information Science and Technology, 29(4): 8-11.

[21] Chandler, A. & Westbrooks, E. L. (2002). Distributing non-MARC metadata: The CUGIR metadata sharing project. Library Collections, Acquisitions, & Technical Services, 26(3): 207-217.

[22] Herold, P., Turner, T. P. & Gale, T. D. (1999?). Final Project Report: Cornell University Geospatial Information Repository (CUGIR).

[23] MODS The Metadata Object Description Schema: The Official Website (2003). Retrieved Feburary 12 2003, from The Library of Congress Network Development and MARC Standards Office Website: http://www.loc.gov/standards/mods/.

[24] Guenther, R. (2003). MODS: The Metadata Object Description Schema. Retrieved April 18 2003, from Portal: Libraries and the Academy Online Journal: http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v003/3.1guenther.html.

[25] METS Metadata Encoding Transmission Standard: Official Web Site (2002). Retrieved May 16 2003, from http://www.loc.gov/standards/mets/.
[26] Guenther, R. & McCallum, S. (Dec 2002/Jan 2003). New Metadata Standards for Digital Resources: MODS and METS. Bulletin of the American Society for Information Science and Technology, 29(2): 12-17.
[27] Kunze, J. A. (2001). A Metadata Kernel for Electronic Permanence. Retrieved March 15 2001, from Proceedings International Conference on Dublin Core and Metadata Applications, Tokyo, Japan Available Online: http://www.nii.ac.jp/dc2001/proceedings/product/paper-27.pdf.
[28] Nelson, M. (2000). Buckets: Smart Objects for Digital Libraries. Unpublished Ph.D., Old Dominion, Norfolk.
[29] Nelson, M. L. (2001). Smart Objects and Open Archives. Retrieved January 15 2002, from D-Lib Magazine 7(2): http://www.dlib.org/dlib/february01/nelson/02nelson.html.
[30] Nelson, M. L. (2003). Smart Objects, Dumb Archives: A User-Centric, Layered Digital Library Framework. Retrieved June 30 2001, from D-Lib Magazine 5(3): http://www.dlib.org/dlib/march99/maly/03maly.html.
[31] The author and Adam Chandler were awarded approximately $3000.00 by the Cornell University Libraries Internal Grant Competition, 2000/2001to implement the "Enhancing Access to Cornell University Geospatial Information Repository (CUGIR): Federal Geospatial Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) to MARC/Dublin Core Conversion Project." The title of the project does not indicate that one of the biggest components of this project was the SODA Model. The costs associated with maintaining SODA and multiple metadata schemas have been absorbed in the CUGIR metadata workflow. Student workers in Technical Services perform the majority of tasks associated with the conversion and updating of metadata records and the creation of the buckets.
[32] Ostensen, O. (2001). Expanding Agenda of Geographic Information Standards. Retrieved May 2 2003, from ISO http://www.iso.ch/iso/en/commcentre/pdf/geographic0107.pdf.
[33] OGC: The Open GIS Consortium, I. (2003). About The Open GIS Consortium, Inc. Retrieved May 2 2003, from Open GIS Consortium, Inc. Website: http://www.opengis.org/ogcAbout.htm.