The Metadata Architecture for Data Management in Web-based Choropleth Maps

Yu Deng
yuzi@cs.umd.edu
Department of Computer Science
University of Maryland
College Park, MD 20742

February 8, 2002

Abstract Metadata is used increasingly to improve both the availability and the quality of the information delivered. Professionals primarily view metadata as a way to analyze and structure the underlying data. However, the use of metadata has additional potential benefits. Metadata has tight connection with “upper data”, data viewed and used by end users. While metadata tailors the user interface of our project, user requirements play an important role in the effectiveness of constructing our metadata. This paper proposes a metadata architecture based on both information integration and user needs, since existing metadata theories and tools do not provide an efficient implementation support. Adopting a common user interface for browsing and querying georeferential data on the Web and building a package of metadata standards for Mapping Information (mapping between information meaningful to users and the raw data), will not only lower the implementation cost, but also enhance the quality of metadata and its architecture. This article offers guidelines and anticipates the future of metadata development.

Keywords: metadata, metadata architecture, choropleth map, dynamic query, standard

1 INTRODUCTION

Vast collections of statistical data from numerous government agencies play an increasingly critical role in the decision-making processes of institutions and organizations from commerce and education to research and healthcare. In addition to publishing the data on CDs and other medias, those agencies are shifting toward making the information available on the Web following the explosive growth of the Internet and its users in recent years. Large volumes of government data can now be widely accessed interactively.

Dramatic opportunities exist to develop direct manipulation interfaces, end-user programming tools, dynamic queries to perform information search in large databases, and information visualization to support network database browsing [15]. Choropleth maps are a commonly used visualization method of displaying geo-spatial data. They are very useful for showing phenomena with distinctive distributions, for revealing geographical patterns, and for discovering outliers (extreme values of variables with unremarkable distributions), etc. [21].

In georeferential information systems, the emerging application of metadata can improve the availability as well as the quality of the information delivered. The growing popularity of Internet-based data servers has accelerated this trend even further [7]. There is a major demand for geographical information and affiliated management tools. Facilitated by the phenomenal expansion of the Internet, government agencies and institutions such as the Federal Geographic Data Committee (FGDC), the US Bureau of Census, and the National Park Service, are shifting toward providing web-based choropleth map services for public access.

This paper discusses the technologies of metadata in geo-spatial information systems, specifically, our understanding of metadata and metadata architecture embodied in our approach to provide web-based choropleth map services. Section 2 of this paper introduces our project, Web-based Dynamaps. Section 3 examines our understanding of metadata versus other existing definitions. Section 4 describes the metadata architecture employed in our project based on the comparison with other common architectures. Section 5 introduces such metadata standards as the Dublin Core and standard of Federal Geographic Data Committee. Then, the metadata standard for Mapping Information and guidelines for data providers is proposed. Section 6 presents findings and conclusion.

2 WEB-BASED DYNAMAPS

A project to create a Web-based version of Dynamaps [6] for US Bureau of Census establishes the framework for this paper. Dynamaps are a generalized map-based information visualization tool for dynamic queries and brushing on choropleth maps. Users can use color-coding to show a variable on each geographic region, and then filter out areas that do not meet the desired criteria. In addition, a scatter-plot view and a details-on-demand window support overviews and specific fact-finding. The non-Web-based version has been implemented using Visual Basic, with a commercial GIS display engine (ESRI) [6]. It was designed by the Human Computer Interaction Laboratory at University of Maryland College Park (http://www.cs.umd.edu/hcil/).

Following the trend of providing Web-based choropleth map services for geo-spatial data, we proposed and implemented a prototype to bring Dynamaps on the Web. Our initial intention was to dynamically query and browse the statistical data using map filters since such use of proximity coding, plus color coding, size coding, animated presentations, and user-controlled selections enable users to explore large information spaces rapidly and reliably [1].

Our architecture for implementing the Web-based Dynamaps was composed of the Data Management component and the Map Service component. The Map Service component is a Web-based choropleth map service for browsing and querying georeferential data. The Data Management component locates and accesses the raw data across multiple repositories. The core of the Data Management component is metadata, which describes the available statistical data sets. The metadata is constantly updated as the raw data changes. The metadata technology enables the administrator to dynamically adjust the data sets by simply updating data entries in the metadata database.

The result of our project is promising. The Web-based Dynamaps using metadata achieve data scalability while supporting such user-oriented functionalities as dynamic query by scrolling sliders, zooming to maps at more detailed levels, and brushing across the choropleth maps.

3 METADATA

3.1 Need for metadata

As massive quantities of data are being produced, stockpiled, accessed or transformed daily in digital repositories distributed across multiple sites, how to efficiently and effectively organize those data to satisfy the users' diverse needs is as essential as the contents of the data. The underlying solution should address such issues as improving accessibility and availability, maintaining and documenting interrelationships among multiple data objects, and preserving data integrity through successive system upgrades. Nowadays, information professionals have designed various retrieval tools to automatically or semi-automatically locate useful data. Metadata forms the foundation of these tools by serving as a condensed representation of the underlying data. As such, it supports browsing, navigation, and content-oriented indexing. Metadata manages the history of changes made to the data. Furthermore, it supervises existing data holdings, unifies naming schemes, and records relationships among different data items and data sets.

In practice, metadata technology is increasingly being integrated into commercial GIS. Most commercial systems have always maintained certain basic metadata on the objects to be administered. ARC INFO, for example, generates and maintains metadata on the spatial registration, projection, and tolerances of a coverage or grid [7].

For a geographic information system, while the data are being accessed by numerous parties and for a spectrum of purposes, the system should provide an integrated view on individual geographic data sets. One key task is providing the end-users as well as the system administration with descriptive information on the data contents in order to facilitate transparent integrated access to diverse information sources. The tasks required in building such an application are closely related to the integration of heterogeneous databases. The approaches taken from this field, involves managing a significant amount of metadata for global query decomposition, global transaction management, schema integration, and management of federated information systems [25].

3.2 Definition of metadata

In addition to the simple definition – structural data about data, metadata has been characterized in definitions that are more formal.

- Metadata is descriptive information about an object or resource whether it be physical or electronic. While the term “metadata” is relatively new, the underlying concepts behind metadata have been in use for as long as collections of information have been organized. For example, library card catalogs represent a well-established type of metadata that has served as collection management and resource discovery tools for decades. Metadata can be generated either "by hand" or derived automatically using software [18].
- Metadata describes the content, quality, condition, and other characteristics of data. Metadata helps a person to locate and understand data. It includes information needed to determine the sets of data that exist for a geographic location, information needed to determine if a set of data meets a specific need, information needed to acquire an identified set of data, and information needed to process and use a set of data. The exact order in which data elements are evaluated, and the relative importance of data elements, will not be the same for all users [19].

A simple example of metadata:
(from http://www.ukoln.ac.uk/metadata/presentations/tlig-1998-03-31/tlig/sld006.htm)
<Meta name = “keywords” CONTENT = “national centre, network information support, library community, awareness, research, information services, public library networking, bibliographic management, distributed library systems, metadata, resource discovery, conferences, lectures, workshops”>

Although the concepts of metadata have been widely accepted and applied in various fields, many current metadata applications are not developed with the emphasis on usability. In our project, however, such emphasis is imperative.

3.2.1 Metadata in Web-based Dynamaps

The metadata in our application area should contain the following characteristics:

- The metadata architecture should be user-oriented, specifically tailored for geo-spatial data users. The tools designed to browse and query the large datasets from georeferential information systems, place the users of all levels at the center of interacting with the application interface features. Thus, the metadata architecture should emphasize the role of users in its design.
- The metadata should allow transparent data integration from different sources. In georeferential information systems, data is organized according to a wide variety of data models. Metadata can help to overcome these heterogeneities by specifying the platforms on which a given data item is located.
- The ever evolving quantities, conditions, and other characteristics of the data sets might render the initial blueprint of the application user interfaces obsolete in a short period of time. The metadata architecture should be designed to reduce such impacts. Although user interfaces should be built independent of the underlying raw data, the robustness of the interfaces would be jeopardized if the dynamic nature of the data were ignored.

3.3 Generate metadata

In the case of structured databases, the norm is to use schema descriptions and associated information (such as database statistics) as metadata. In the case of unstructured textual data and information retrieval, metadata is generally limited to indexes and textual descriptions of data. Metadata in such cases provides a suitable basis for building the higher forms of information [25].

3.3.1 Methods for generating metadata

Metadata is commonly generated via three methods, namely, Analysis of Raw Material, Semi-automatic Augmentation, and Processing with Implicit Metadata Generation [25].

- Analysis of Raw Material

In many cases, media objects are analyzed and metadata is generated according to the focus of the analysis.

- Semi-automatic Augmentation

Semi-automatic augmentation of media results in additional meta-information, which cannot be derived from the raw material as such.

- Processing with Implicit Metadata Generation

Metadata can be generated implicitly when creating the raw media data.

Metadata has to be updated according to changes in raw data. This may cause a new full cycle of metadata generation in order to substitute the old existing metadata on a medium. Furthermore, metadata can be updated directly without any modification of the raw material. A simple example is the correction of metadata according to the changes of semantic knowledge used for constructing the metadata.

3.3.2 Some experience for generating metadata

Generating the metadata can easily be a tedious task although using automatic tools may alleviate some of the pain. The task is more daunting when attempting to generate a huge volume of metadata without knowing the data, its usage, its background knowledge, and its accuracy, etc.

Before generating the metadata, it is necessary to review all the relevant documentation about the data. It would be far more effective to let individuals proficient with the data to tackle the task if possible. Any data anomalies should trigger necessary measures as soon as they're noted. Erroneous metadata can be fatal to the integrity of the data.

3.4 Types and usages of metadata

Metadata can be categorized into three types [25]:

- Media type-specific metadata. Media types induce specific kinds of metadata, e.g., texture of images. The more specific a media type, the more specific the associated metadata attributes.

- Media processing-specific metadata. Metadata can describe functions designed to process specific media. Another important type of media processing-specific metadata is information related to media processing performance, which can be used to measure and consequently achieve desirable system performance. Similarly, meta-information about the interoperability of system components is essential to deliver the proper application functionality.
- Content-specific metadata. This kind of metadata is solely derived from the content represented by the objects.

Despite the variance in metadata types, the most usages for metadata include querying, retrieval, navigation, and browsing. Furthermore, metadata can support such queries as the life cycle of certain raw data, which cannot be answered by processing raw data alone.

3.5 Some noticeable problems about metadata

Some problems should be noticed [17]:

- The costs of not creating metadata are much bigger than the costs of creating since the loss of information with staff changes, data redundancy, data conflicts, liability, misapplications, and decisions are all based upon poorly documented data.
- Don’t try to cover all of the data resources with a single metadata record. A good rule of thumb is to consider how the data resource is used – as a component of a broader data set or as a stand-alone product that may be mixed and matched with a range of other data resources.
- Human review matters. The whole process of creating metadata should not rely solely on automated tools.
- Assessments of consistency, accuracy, completeness, and precision about data are quite important. The methods for controlling the data quality include field checks, cross-referencing, and statistical analyses, etc.
- Metadata should be recorded throughout the life of a data set, from planning (entities and attributes), to digitizing (abscissa/ordinate resolution), to analysis (processing history), through publication (publication date). Organizations and agencies are encouraged to develop operational procedures that institutionalize metadata production and maintenance, and make metadata a key component of their data development and management process.

4 METADATA ARCHITECTURE

4.1 Introduction to metadata architecture

Originally, the use of metadata has been associated to schema descriptions maintained in database management system catalogues or either repository entries, providing information about both stored data and business processes associated with them. More recently, there is a major agreement that the use of metadata constitutes the main factor to promote integration and information exchange amongst heterogeneous digital sources. Many projects under way admit that it is unlikely that some unique metadata format or standard will be universally used. They recognize the need for a higher-level container architecture that can accommodate different metadata standards already in use, establishing general frameworks where many different initiatives could coexist [22].

4.1.1 Warwick Framework

The Warwick Framework [23] provides the conceptual structure for aggregating physically distinct metadata sets. It is mainly for aggregating multiple sets of metadata.
The main principles of this architecture are:

- Metadata takes a variety of forms, both specialized and general. These forms include descriptive cataloging, terms and conditions (this is metadata that describes the conditions for use of an object), administrative data (this is metadata related to the management of an object in a particular server or repository), content ratings (this is a description of attributes of an object within a multidimensional scaled rating scheme), provenance (this is data defining source of origin of some content object), linkage or relationship data (this is data about the relationship of a content object to other objects) and structural data (this is data defining the logical components of complex or compound objects and how to access those components).
- The architecture must be sufficiently flexible to incorporate new semantics without requiring a rewrite of existing metadata sets.
- Different communities will propose, design, and be responsible for different types of metadata.
- Different metadata sets are used by and may be restricted to distinct communities of users and agents.
- Metadata and data have similar behaviors and characteristics. What may appear to be metadata in one context may look very much like data in another.

The Warwick Framework has two fundamental components: container and package. A container is the unit for aggregating the typed metadata sets, which are known as packages. The only operation defined for a container is one that returns a sequence of packages in the container. Any encoding for a container must allow the recipient of the container to skip over unknown packages within the container. Packages are of three types:

- Metadata set (These are packages that contain actual metadata)
- Indirect (This is a package that is an indirect reference to another object in the information infrastructure)
- Container (This is a package that is itself a container)

The figure below shows a simple example of a Warwick Framework container. The container in this example contains three logical packages of metadata. The first two, a Dublin Core (introduced in section 5.1.1) record and a MARC (introduced in section 5.1.4) record, are contained within the container as a pair of packages. The third metadata set, which defines the terms and conditions for access to the content object, is referenced indirectly via a URI (“Uniform Resource Identifier, a compact string of characters for identifying an abstract or physical resource” [28]) in the container.

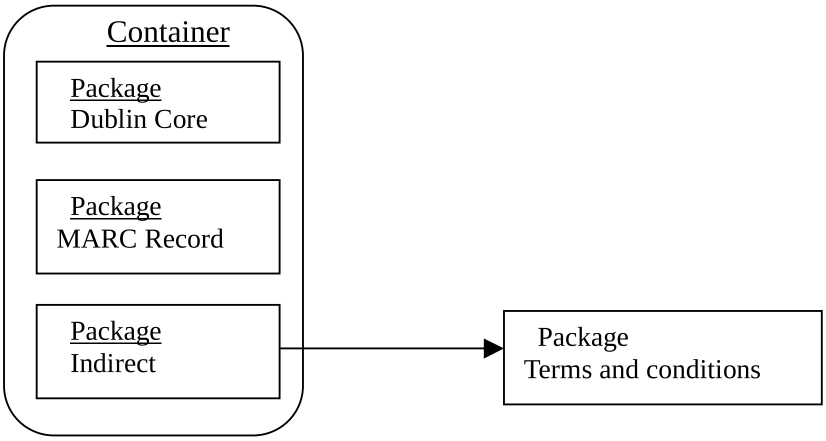


Figure 1. An example of Warwick Framework container

One of the challenges presented by this model is how relationships among packages may be semantically and operationally defined inside containers. Relationship types specify a context for each package, specifying if it plays a data or metadata role. A new component, named Catalog, has been introduced in the Warwick Framework to describe relationships among packages in digital objects [22].

4.1.2 Metadata architecture to represent electronic documents

A metadata architecture to represent electronic documents on the Web is proposed by Moura, Campos and Barreto [22]. In the conceptual model of this architeture, an electronic document is represented as a hierarchy of interrelated levels. At each level, different metadata types would be needed to represent and describe the document.

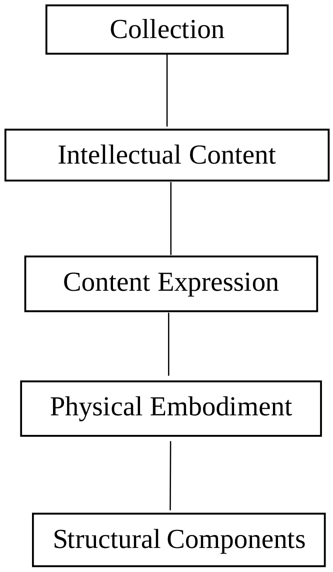


Figure 2. Conceptual model of the architecture

In figure 2, Collection represents information resource, such as a document. Intellectual Content expresses someone's or some organization's ideas and opinions about a subject matter. The title of a paper can be the intellectual content. Content Expression specifies how this content is organized. A document content expression may be classified into categories such as: a technical report, an article, a manual, a newspaper, etc. At next level, Physical Embodiment is used as a mean to disseminate this expression. The format of a paper, for example, is a kind of physical embodiment. At the lowestest level, Structural Components represent a form on which its physical embodiments may be segmented. For instance, a word document can be divided into pages, paragraphs, sentences, etc.

One implementation of this architecture is to use simple components named digital objects, which are proposed in Warwick Framework and mainly composed of containers and packages illustrated in figure 1. Digital objects are grouped together into a hierarchical recursive representation for the conceptual model. In this hierarchy the root of the digital object class represents the superclass of all digital objects. Each one implements a level of the document conceptual model. It has the following common attributes:

- ObjectType: describes the type of the digital object, such as ContextualObject, StructuralObject.
- ObjectName: a string used for object identification.
- MetadataContainer: contains or references data that are not part of resource content. It associates metadata packages to the correspondent digital object.
- DataContainer: contains or references data that are part of the resource content. It allows the representation of a document structure according to different levels of granularity.

4.1.3 RDF

The Resource Description Framework (RDF) is a W3C proposed infrastructure that enables the encoding, exchange and reuse of structured metadata. According to the W3C RDF FAQ, “RDF emphasizes facilities to enable automated processing of Web resources”. RDF metadata can be used in a variety of application areas; for example: in resource discovery to provide better search engine capabilities; in cataloging for describing the content and content relationships available at a particular Web site, page, or digital library; by intelligent software agents to facilitate knowledge sharing and exchange; in content rating; in describing collections of pages that represent a single logical “document”; for describing intellectual property rights of Web pages, and in many others [2].

RDF is an application of XML that imposes needed structural constraints to provide unambiguous methods of expressing semantics. RDF additionally provides a means for publishing both human-readable and machine-processable vocabularies designed to encourage the reuse and extension of metadata semantics among disparate information communities. This infrastructure enables metadata interoperability through the design of mechanisms that support common conventions of semantics, syntax, and structure.

RDF supports the use of conventions that will facilitate modular interoperability among separate metadata element sets as well as the combination of distributed attributes.

- RDF Data Model
  RDF provides a model for describing resources, which have properties (attributes or characteristics). The properties associated with resources are identified by property-types, and property-types have corresponding values. Using a triadic model of resources, property-types and corresponding values, RDF attempts to provide an unambiguous method of expressing semantics in a machine-readable encoding. The unambiguous identification of resources provides for the reuse of explicit, descriptive information.
- RDF Syntax
  RDF defines a simple, yet powerful model for describing resources. It provides the ability for resource description communities to define semantics. However, it is important to disambiguate these semantics among communities. To prevent ambiguity, RDF uniquely identifies property-types by using the XML namespace mechanism. XML namespaces provide a method for unambiguously identifying the semantics and conventions governing the particular use of property-types by uniquely identifying the governing authority of the vocabulary. Actually, the structural constraints RDF imposes to support the consistent encoding and exchange of standardized metadata provides for the interchangeability of separate packages of metadata defined by different resource description communities.
- RDF Schema
  RDF Schemas are used to declare vocabularies, the sets of semantics property-types defined by a particular community. The XML namespace mechanism serves to identify RDF Schemas. It is anticipated, however, that the ability to formalize human-readable and machine-processable vocabularies will encourage the exchange,

> use, and extension of metadata vocabularies among disparate information communities. RDF schemas are being designed to provide this type of formalization.

RDF is quite different from the above two architectures. It is based on the triadic model of resources, property-types and corresponding values and has tight connection with XML. The following is an example of RDF for describing a Web page:
(this example is from http://www.ibiblio.org/xml/slides/sd2001east/fundamentals/42.html)

```
<rdf:RDF
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:dc="http://purl.org/DC/>
     <rdf:Description about="http://www.ibiblio.org/xml/>
          <dc:CREATOR>Elliotte Rusty Harold</dc:CREATOR>
          <dc:TITLE>Cafe con Leche</dc:TITLE>
     </rdf:Description>
</rdf:RDF>
```

4.2 Metadata architecture for Web-based Dynamaps

All of the common architectures discussed above emphasize the data description and information integration. The Warwick Framework is mainly for aggregating multiple sets of metadata. The main purpose of Moura's architecture is to allow the description of associations and collections of electronic documents from different information categories, and to provide a solution for integrating the initiatives, such as Dublin Core [20], with a vast quantity of descriptive information [22].

These architectures may encounter shortcomings in our project for the following reasons:

- Since Web users may not be experts on the raw data sets, mapping between the raw data and the information meaningful to the users is required. Nonetheless, no metadata architecture mentioned above can address how such mapping can be achieved.
- Administrative users may be confronted with unfamiliar data in existing or updated systems. They may also need to trace the origin of certain data while administrating the architecture. Without providing historical mapping information of data, whether Warwick Framework or Moura's architecture won't assist such users in attaining the desired results.
- The raw data in the Web-based Dynamaps project are statistical data sets. The data management component for those data sets in our approach should produce information with specific formats delivered to the client side for the map service. The current architectures fail to produce that information.

4.2.1 Details of the metadata architecture for Web-based Dynamaps

Unlike the metadata element in HTML, which lets the author specify the information about a document such as author name and creation date, the Web-based Dynamaps metadata provides semantics about the data sets and data attributes. The metadata architecture is a framework that integrates multiple levels of users, the raw datasets, the metadata, and various information retrieval methods.

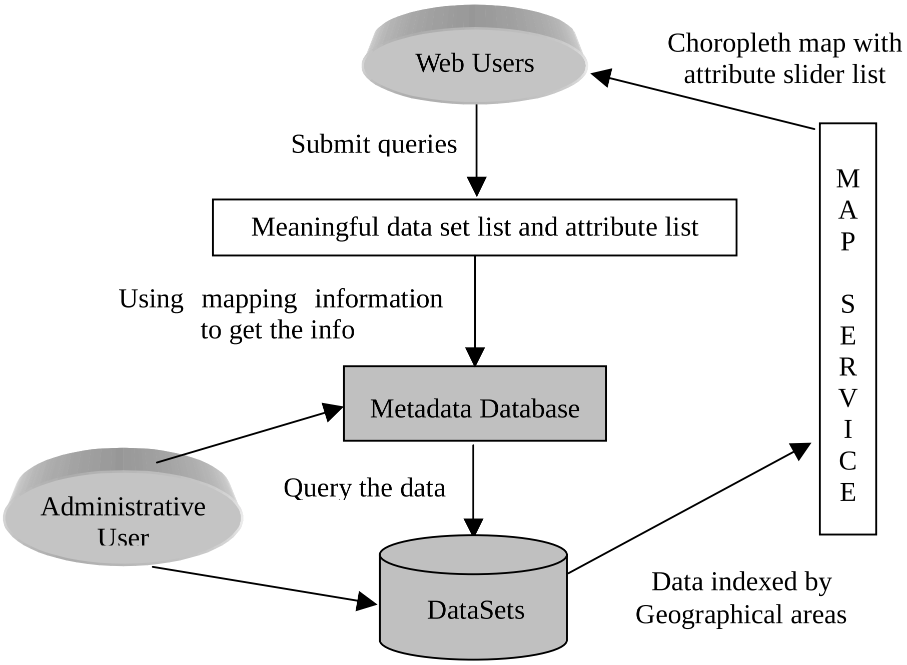
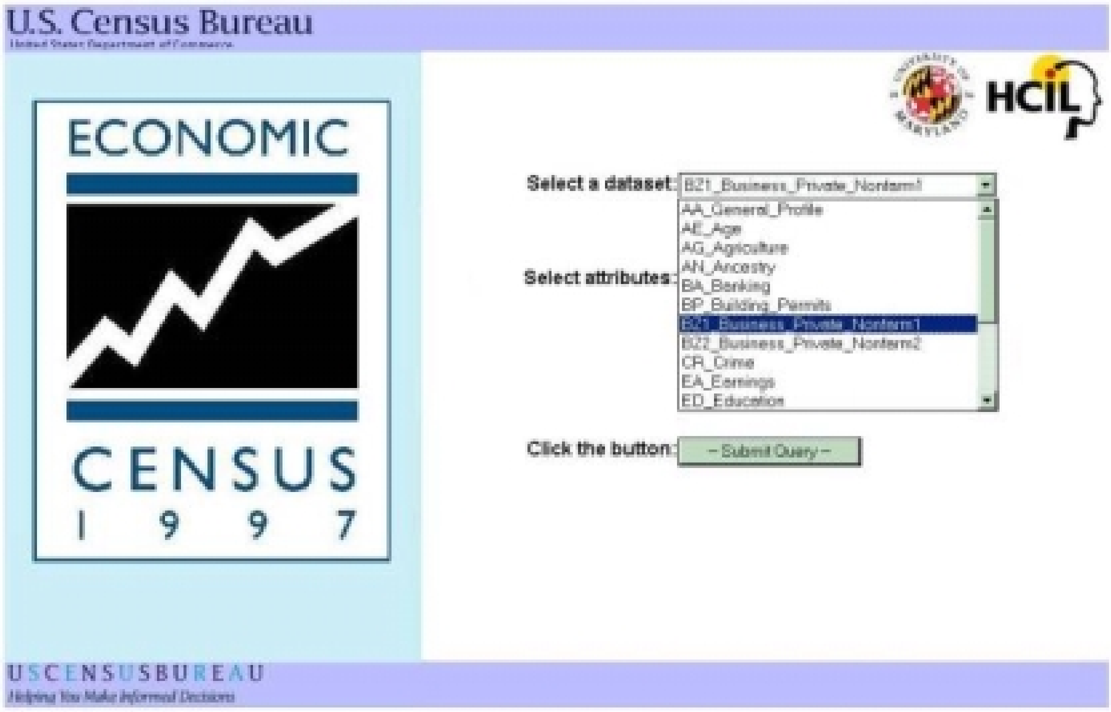

Web Users
Choropleth map with attribute slider list
Submit queries
Meaningful data set list and attribute list
Using mapping information to get the info
Metadata Database
MAP SERVICE
Administrative User
Query the data
DataSets
Data indexed by Geographical areas

Figure 3. Metadata architecture for Dynamaps

In this framework, the two levels of users are: Web users and administrative users. The interface for Web users is the set of JSP pages that present the meaningful data sets and attributes. The Web users can choose the data sets as well as attributes, and then submit the query.



Select a dataset: BZ1_Business_Private_Nonfarm1

Select attributes:
AA_General_Profile
AE_Age
AG_Agriculture
AN_Ancestry
BA_Banking
BP_Building_Permits
BZ1_Business_Private_Nonfarm1
BZ2_Business_Private_Nonfarm2
CR_Crime
EA_Earnings
ED_Education

Click the button: - Submit Query -

Figure 4. A JSP page for the data sets of Econ97

To communicate between the raw data sets and the information meaningful to the Web users would require an interface mechanism. In the case of the web-based Dynamaps, the interface is the metadata, which would be managed by the administrative users. Through the metadata, both the administrators and Web users are able to find out what the data sets are available and the descriptions of those data sets in details. The operations of the administrative users include insertion, deletion of data sets and making appropriate updates in metadata database. The administrative users must go through the following steps whenever new data sets are inserted:

- Generate the metadata database by using our metadata tool.
- Double-check the correctness and integrity of the data sets and metadata.
- Publish the data sets and attributes on the Web.

With the map service, Web users can explore particular trends/patterns in the data and complete some specific scenario tasks [6], e.g. discovering which state has the largest population. They can utilize the map service by following the steps below:

1. Choose the data set about population.
2. Choose the attributes about the state population in a period of time.
3. Submit the query.
4. Retrieve the choropleth map with the desired data and do filtering on the map to obtain the desired information.

After the users submit the query in step 3 above, the system must complete the following tasks before step 4:

- Find the real names of the attributes from the metadata database.
- Find the table names of the data set from the metadata database.
- Query the data sets to get the desired data.
- Order the data by geo-spatial area and send it to the component of Map Service in specific format.

The following example illustrates the administrative users' actions during a metadata configuration.

Suppose there are two new data sets to be inserted: Costat01.dbf and Costat02.dbf, which are two data files from the CD of USA Counties 1998 prepared by the Bureau of the Census. They provide the statistical data about age. In addition, the attributes in each data set are known.

Each record in these two data sets begins with the eight attributes describing geographic information.

FIELD	FIELD NAME	TYPE	WIDTH	DESCRIPTION
1	SEG	Character	2	File number (e.g. COSTATxx.DBF)
2	STCOU	Character	5	FIPS state and county code (as of 1/1/92)
3	ST	Character	2	FIPS state code
4	COU	Character	3	FIPS county code (as of 1/1/92)
5	SUMLEV	Character	1	Geographic level: 0 = U.S. 1 = State 2 = CMSA/PMSA County (inside New England) 3 = MSA County (Outside New England) 4 = NECMA County 5 = County not in metro area
6	METRO	Character	4	CMSA/MSA/NECMA code (as of 6/30/98)
7	PMSA	Character	4	PMSA code (as of 6/30/98)
8	AREA NAME	Character	36	Name of area

Table 1. First eight attributes of the data sets
(MSA = metropolitan statistical area; CMSA = consolidated MSA; NECMA = New England county MA; PMSA = primary MSA; FIPS = Federal Information Processing Standards)

After the geographical attributes are the statistical attributes.

Every statistical attribute name is an 8-character string. The first two characters are the two-letter table abbreviation. For example, “AE” is the abbreviation of Age table. The third through fifth characters of the 8-character string is a 3-digit code. These three digits uniquely identify each data item in a table. For example, the definition of code “007” represents “RESIDENT POPULATION (100%)”(U.S. Bureau of the Census, Census of Population and Housing, Summary Tape Files 1C and 3C.). Characters 6 and 7 of the statistical attribute indicate the year of the data. And the 8th character, an "F" or "D", is used to differentiate the flag field for the attribute (F) from the actual data (D). So, an attribute “AE00780D” means the resident population (100%) in 1980 for all age ranges and the source is Bureau of the Census. Since Costat01.dbf and Costat02.dbf are data files about age, the first two characters of all the statistical attributes are “AE”.

The first problem encountered is that the names of these two sets mean nothing to the average Web users. The second problem is how to access the real data sets once the system receives the queries in the form of meaningful names submitted by the Web users.

To fix these two problems, a mapping mechanism should be created between the original names and the meaningful names and save enough details in the metadata. Using the mapping mechanism, the system can locate the data sets and attributes by the meaningful names and present the meaningful names corresponding to the available data sets and attributes.

In this case, the metadata can be defined as the following (please see figure 5).

‘DataSetId’ is the primary key in the table of data sets and the foreign key in the table of attributes. ‘DataSetName’ is the meaningful name for a data set. And ‘DataSetFile’ is the path to find the real data set.

In the table of attributes, ‘AttributeName’ is the meaningful name for an attribute; ‘ColumnName’ is the name for the same attribute in the data set file; ‘ColumnType’ is the data type for the attribute, such as string, number, etc.

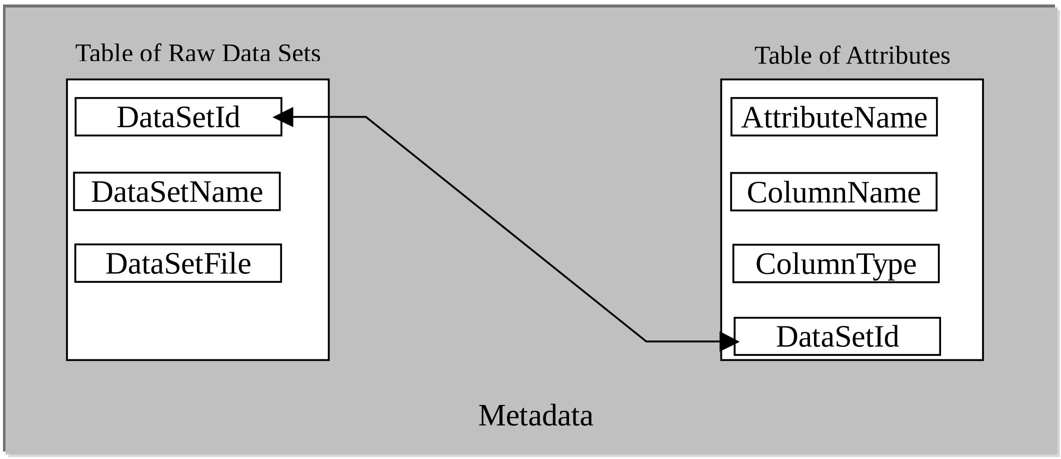


Figure 5. Sample structure of metadata

For data set Costat01.dbf and Costat02.dbf, two new records will be added to the table of data sets:

DataSetId	DataSetName	DataSetFile
1	Age	Costat01
2	Age	Costat02

The table of attributes contains records describing the attributes of the raw data sets like the following:

AttributeName	ColumnName	ColumnType	DataSetId
Resident population for all age ranges in 1980 (100%)	AE00780D	Number	1
Population of people under 5 years in 1980 (100%)	AE02080D	Number	1
Population of people from 5 to 14 years in 1980 (100%)	AE03080D	Number	1
Population of people under 1 year in 1990 (Sample)	AE51090D	Number	2
Population of people from 1 to 4 years in 1990 (Sample)	AE52090D	Number	2
Population of people from 5 to 9 years in 1980 (Sample)	AE61080D	Number	2
…	…	…	…

4.2.2 Characteristics

The characteristics of our metadata architecture include:

- Dynamic data loading: The JSP pages are dynamically generated from the metadata. The use of metadata ensures that the Web users can always view the updated information about what data sets and attributes are available. Whenever there are changes in metadata, the JSP pages can show them immediately after reloading.
- Good scalability: Our metadata architecture helps to achieve good scalability of the system. When changes are made to the data sets or their organizations, the administrative users should only need to update the content of the metadata reflecting the changes. There is no need to update the codes for querying the metadata and for querying the raw data.
- Multi-levels for users: the architecture holds two levels of users, Web users and administrative users, as well as two levels of user interfaces, Web user interface (choropleth map) and the administrative user interface.

4.2.3 Problems and solutions

The data sets from the different sources may have different formats or rules. This will result in difficulties to understand the semantics of data for the administrative users. One solution is to define the semantics of data in the metadata or convert the metadata of the new data sets to our metadata format. So the main program will not change upon the change of data sets.

However, this solution is semantics dependent. There might still be problems when retrieving information from the real data sets with different formats or rules. For instance, a unique record may or may not be identified by a State Id and a County Id depending on the datasets. Another example would be the value of certain attributes. In some data sets, if there were no value for an attribute, it would just be empty. While in other data sets, this attribute may be “0” for a record. Although the different situations can be recorded in metadata and handled with different methods, the costs are quite high when programs must be written to retrieve the data. Why not set the standard for the concerned parties?

The adoption of such a standard helps lower the cost of developing and maintaining commercial georeferential information systems. In the next section, several existing standards and our proposed guidelines for standards will be discussed.

5 METADATA STANDARD

Using metadata in GIS becomes essential, because of the variety and complexity of data types involved as well as their diverse interrelationships. To emphasize the significance of that factor and the importance of using metadata in GIS, specific standard descriptions for georeferential data have been developed [24]. Standards are an important means to achieve common representation schemes and interoperability of systems, and hence can play a pivotal role in exploiting metadata [7].

The key benefits of developing standards are:

- Improve the standardization of policies and data.

- Encourage the consistency in data generation and use.
- Reduce the time and control the costs for processing the data.
- Provide better management of the data.
- Reuse previously generated data
- Reduce redundancy of information.
- Retain scalability and flexibility when exchanging the data.

5.1 Review of Metadata Standards

5.1.1 Dublin Core

The Dublin Core Metadata Workshop Series is an ongoing effort to formulate an alternative description standard for networked objects. The goal is to develop a core element set that provides adequate data for Web resource discovery and is simple for authors and content managers to create and maintain.

The Dublin Core effort has produced two significant results:
1. The specification of the names and semantics of fifteen core descriptive metadata elements (the so-called Dublin Core).

2. The specification of a broader container framework for the Dublin Core and metadata in general (the so-called Warwick Framework, which has been described above).

The Dublin Core Metadata Element Set is a set of 15 descriptive semantic definitions. It represents a core set of elements likely to be useful across a broad range of vertical industries and disciplines of study [18].

The Set was also created to provide a core set of elements that could be shared across disciplines or within any type of organization needing to organize and classify information [18].

The scope of the Dublin Core was specifically designed to provide a metadata vocabulary of "core" properties able to provide basic descriptive information about any kind of resource, regardless of the media format, area of specialization or cultural origin. It is important that a semantic model used for resource discovery is not dependent on the medium of the resource it means to describe [18].

The Dublin Core metadata vocabulary is the result of many years of collaborative research to determine a common set of properties universal for describing any type of resource. The use of a standardized general classifications system also enables metadata of such collections to be combined and for knowledge contained within each collection to be shared [18].

The fifteen Dublin Core elements are:

Element Name	Definition
Title	A name given to the resource
Creator	An entity primarily responsible for making the content of the resource
Subject and Keywords	The topic of the content of the resource
Description	An account of the content of the resource
Publisher	An entity responsible for making the resource available
Contributor	An entity responsible for making contributions to the content of the resource
Date	A date associated with an event in the life cycle of the resource
Resource Type	The nature or genre of the content of the resource
Format	The physical or digital manifestation of the resource
Resource Identifier	An unambiguous reference to the resource within a given context
Source	A Reference to a resource from which the present resource is derived
Language	A language of the intellectual content of the resource
Relation	A reference to a related resource
Coverage	The extent or scope of the content of the resource
Rights Management	Information about rights held in and over the resource

Table 2. Dublin Core elements

Although it may improve structured access to information on the Internet and promote interoperability among disparate description models [7], Dublin Core is still not enough for our project since its set of elements is too general and too small.

5.1.2 FGDC

The Federal Geographic Data Committee has drawn up the Content Standard for Digital Geospatial Metadata [14]. This standard establishes the names of data elements and compound elements (groups of data elements) to be used for providing a common set of terminology and definitions for the documentation of digital geospatial data, the definitions of these compound elements and data elements, and information about the values that are to be provided for the data elements.

This standard is intended to support the collection and processing of geospatial metadata and intended to be useable by all levels of government and the private sector.

The content of this standard consists of seven parts:

- Identification Information: basic information about the data set.
- Data Quality Information: a general assessment of the quality of the data set, such as the accuracy of information.
- Spatial Data Organization Information: the mechanism used to represent spatial information in the data set, such as Point and Vector Object Information.
- Spatial Reference Information: the description of the reference frame for, and the means to encode, coordinates in the data set, such as the definition of Horizontal Coordinate System.
- Entity and Attribute Information: details about the information content of the data set, including the entity types, their attributes, and the domains from which attribute values may be assigned.
- Distribution Information: information about the distributor of and options for obtaining the data set.
- Metadata Reference Information: information on the correctness of the metadata information, and the responsible party, such as the version of the metadata standard.

5.1.3 SDTS

The Spatial Data Transfer Standard(SDTS) is a Federal Information Processing Standard to facilitate the online exchange of spatial data. It is not an exchange format. Actually, it provides guidelines that need to be translated into a native application-specific format before they can be used [7].

While both the SDTS and FGDC Content Standards refer to metadata about spatial data, they have distinctly separate functions. The SDTS is a language for communicating spatial data across different platform without losing any structural or topological information. The FGDC Content Standards, on the other hand, specify the kind of annotative metadata that federal agencies are required to collect on a spatial data set they maintain. The only two sections that both standards have in common concern data quality and the data dictionary information.

5.1.4 MARC

The MARC standard (Machine Readable Catalogue), one of the metadata standards for bibliographic cataloguing, was created in the late sixties in order to help classification services to enable an exchange of catalogue records among them. It has been used in library automation services, as the basis for manipulating library records for display and indexing [24].

5.1.5 TEI

The main purpose of TEI (Text Encoding Initiative) was to define a set of generic rules for representing textual materials in electronic form, allowing resource interchange and reuse. The initial project aimed to develop guidelines to prepare and interchange eletronic texts for scholarly research. However, TEI guidelines are oriented to the description of objects and give no consideration for service descriptions [24].

5.1.6 FIPS

Federal Information Processing Standards (FIPS) codes are issued by the National Institute of Standards and Technology for a variety of geographical entities, including States, counties, metropolitan areas, and places. The objective of the FIPS codes is to improve the transferability of the data resources within the Federal Government and to avoid unnecessary duplication and incompatibilities in the collection, processing, and dissemination of data.

Each State and the District of Columbia is assigned a two-digit FIPS code. The FIPS State code is a sequential numbering, with some gaps, of the alphabetic arrangement of the States and the District of Columbia: Alabama (01) to Wyoming (56). The State codes are presented as one variable on all data files (ST); the State code for the U.S. total is "00". The combination of a FIPS two-digit State code followed by a FIPS three-digit county code provides a unique geographic identifier for each county and equivalent area.

5.1.7 GILS

The main purpose of GILS is to provide a mechanism for locating useful information generated by many government agencies. It was created as an initiative of the US federal government to help people find information resources throughout its many agencies. GILS identifies and describes these resources, supplementing other government and commercial information-dissemination mechanisms. In a broader sense, GILS can be defined as a decentralized collection of locators and associated information services used by the public to find information, either directly or through intermediaries. GILS defines around 70 registered attributes (called GILS core elements) and adopts the ANSI Z39.50 standard protocol to specify how electronic network searches can be expressed and how results are returned. An important aspect of this standard is to ensure interoperability on a semantic level with the many different GILS servers [24].

5.1.8 Z39.50

Z39.50 is a national standard defining a protocol for computer-to-computer information retrieval. Z39.50 makes it possible for a user in one system to search and retrieve information from other computer systems (that have also implemented Z39.50) without knowing the search syntax that is used by those other systems. Z39.50 is an American National Standard that was originally approved by the National Information Standards Organization (NISO) in 1988 [24].

In Z39.50 (version 3), client and server really understood very little of the semantics of the information being searched and retrieved. The responsibility for this was placed primarily on the human user of the client software. Here, the Z39.50 protocol interactions were a much more directly exposed to the user and shared semantics were only used at a mechanical level, for example to agree on the data type of a particular data element. Neither the client nor the server really understood the meaning of the information that was being searched and retrieved [26].

5.2 Suggestion

Although many standards are available, there is still a need to form a specific standard in our application area (Web-based Dynamaps for georeferential data) under the conformity to current common standards to address the issues described in section 4.2.3.

The specific standard should include two sub-standards:

- A common user interface for browsing and querying
- A standard understanding of attribute semantics (semantic standard)

5.2.1 Common User Interface for Georeferential Data

Dynamaps could indeed be the first prototype to employ such a common user interface for visualizing georeferential data on the Web. Using the common user interfaces can yield the following benefits:

- The users only need to learn a single interface that is applicable to all the geographical data sets. The availability of a common user interface eliminates the need to master different interfaces for different databases of different agencies or organizations.
- A common interface could provide a consistent, reliable means of displaying information, which is very convenient for users.

5.2.2 Semantic Standard

The major purpose of the semantic standard is to provide a common set of terminology and definitions for the agencies and organizations to publish georeferential data on the Web.

There could be two steps for creating application-specific metadata:

1. Produce some general metadata according to the common metadata standards, such as the Identification Information in FGDC.
2. Produce the specific metadata in our application area, e.g., the mapping between information meaningful to users and the raw data (Mapping Information).

In the data sets involved in our project, there are two kinds of attributes: data attributes and control attributes. Data attributes describe the statistical information, such as the total population in 1997. Control attributes are the attributes to identify the geographical area or some classification codes, such as county id and NAICS (North American Industry Classification System) code.

Now, let’s discuss the problem in section 4.2.3 again.

The naming rule for data set attributes employed in the Web-based Dynamaps prototype is introduced in section 4.2.1. After our implementation of the prototype, new data sets are received to be inserted. The naming rule of the new data sets differ entirely from the one used in the original data sets. For example, in the new data sets, the column name "ELECTRI" means "Electricity costs" and "EMP" means "Number of employees". It seems that the columns are named by their meanings.

To fix this problem, it is crucial for the data provider to attach the semantics of the attributes with the data sets. On the other hand, a set of attribute names (meaningful names for the Web users) with standard semantics is needed for Dynamaps on the Web. Therefore, no matter what the real column names of the attributes are in a data set, our metadata generator could construct the mapping between the column names and attribute names successfully as long as the data provide uses the standard semantics.

This set of standard names and semantics should be enlarged when new data sets are published. It is necessary to provide the guidelines for enlarging the set.

5.2.3 Guidelines to expand the semantic standard

Since there would always be new data sets made available, the data producers or the user community should continue to establish and expand the standards for the names and the semantics following the guidelines presented below.

- Enlarged names and semantics should be formally documented with the existing set so that the metadata users can refer to these names and semantics.
- Enlarged attributes should not be used to change the name, definition, type, or domain of an existing attribute. In particular, an enlarged attribute cannot be nested under an old attribute.
- The name and definition given to the enlarged attribute should not be the name and definition of any existing attribute.

5.2.4 Guidelines of generating data sets and attached metadata

According to our experiences in handling data sets and metadata, the data providers had better think carefully about the following questions when creating data sets and attached metadata:

- The data providers should adopt a protocol for control or data attributes. When there is no value for data attributes with a specific set of control attribute values in a record, should the values of the data attributes be "0", or empty, or should this record be deleted? This problem is tightly related to the program access to the data sets. If different providers employ different rules, that program might have exceptions when running.
- The data providers should specify which control attributes have been used and attach the specification and other respective documentation on the attributes with the data

sets. Since the large number of control attributes currently available are the keys in deciding a unique record in the data sets. Such specification would help to construct the JSP pages of Dynamaps, which present the users with control attributes and let them query the desired information.

- As discussed above, the Mapping Information should be included in the metadata of the data sets since the naming rules of different data sets might be quite different. It is the key for Dynamaps to access the data sets with distinctive naming rules.
- A data provider should check the specific standards in an application area and the related common standards if its published data is used in that area. The data sets published by the provider are guaranteed to conform to the semantic standards and be understandable by the maximum users. Also, the Mapping Information would be correct under this conformity.

6 CONCLUSION

Without a doubt, metadata will continue to connect end-users and the underlying raw data. As user requirements become ever more diverse and complex, more attentions should be directed toward satisfying their needs when building the metadata. In this paper, Mapping Information, which maps between the underlying raw data and the information meaningful to the Web users, is recommended to be part of the standards for constructing metadata as a way to provide user-oriented information for both the administrative users and the end users. Another finding lies in the fact that information concerning where the data comes from, how it is produced, who publishes such data, and the like, occasionally, outweighs the raw data itself. For users such as researchers, stock analysts and managers, who often require not only data but also the information about that data, such information is contained in the metadata.

Finally, the use of metadata concepts should not be confined to handling the underlying data. Considering the fact that the quality of metadata and its architecture would affect the user interface, I anticipate that one of the directions for metadata development will be to emphasize user interface design.

7 ACKNOWLEDGEMENT

I would like to thank Professor Ben Shneiderman, Founding Director of the Human Computer Interface Lab at the University of Maryland, College Park, for guiding me on this paper and the development of metadata part. I would also like to thank Dr. Evan Golub of the Department of Computer Science at University of Maryland, College Park, for his encouragement and help. I wish to thank Gunjan Dang, graduate student from the Department of Computer Engineering for her tutorials on the original Dynamaps written in Visual Basic. I wish to acknowledge Mr. Rick Ayers of ESRI Inc. and Dr. David Mingay of the Census Bureau, for their generous help when we implemented the prototype. Finally, I would like to thank the other two members of the team for our joyful cooperation during the course of the project: Haixia Zhao, who was responsible for the implementation of Map Service, and Amit Thakkar, who also took part in implementing Map Service. Both of

them are students from the Department of Computer Science at University of Maryland College Park.

REFERENCES

[1] C. Ahlberg and B. Shneiderman, 1994. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *Proc. CHI'94 Conference: Human Factors in Computing Systems, ACM*, New York, NY (1994), 313-321.

[2] Martin Dillon ,2001. Metadata for Web Resources: How Metadata Works on the Web. Library of Congress, January 23, 2001, http://www.loc.gov/catdir/bibcontrol/dillon_paper.html

[3] Putting Information Together: Building Integrated Data Repositories, Summary Report. Center for Technology in Government, 2000, http://www.ctg.albany.edu/resources/htmlrpt/uig_seminar4/bldg_integrated_data_repositories.html

[4] A.M. MacEachren, F. Hardisty, M. Gahegan, M. Wheeler, X. Dai, D. Guo and M. Takatsuka, 2001. Supporting visual integration and analysis of geospatially-referenced statistics through web-deployable, cross-platform tools, Proceedings, dg.o.2001, *2001 National Conference for Digital Government Research*, Los Angeles, CA, May 21-23.

[5] B. Shneiderman, 1999. Dynamic queries, starfield displays, and the path to Spotfire. February 4, 1999.

[6] G. Dang, C. North, B. Shneiderman, 2001. Dynamic Queries and Brushing on Choropleth Maps. *Proc. IEEE Intl. Information Visualization*, 2001.

[7] O. Gnther and A. Voisard, 1998. Metadata in Geographic and Environmental Data Managment in *W. Klas, A. Sheth (eds.) Managing Multimedia Data: Using Metadata to Integrate and Apply Digital Data*. McGraw Hill, 1998.

[8] T. Berners-Lee and D. Connolly, 1995. Hypertext Markup Language - 2.0, MIT/W3C , September 22, 1995. http://www.w3.org/MarkUp/html-spec/html-spec_toc.html

[9] Brand Niemann, 1998, A GIS Starter Kit for Sustainable Development in the Southern Appalachian Area, *Ninth Annual SAMAB Annual Fall Conference*, November 4-6, 1998, Gatlinburg, TN.

[10] Brandow Plewe, 1997. So You want to Build an Online GIS? *GIS World*, Nov. 1997, pp. 58-60. http://kayenta.geog.byu.edu/gisonline/

[11] Federal Geographic Data Committee, http://www.fgdc.gov/, last access: December 2001.

[12] Dublin Core Metadata Initiative, http://dublincore.org/, last access: December 2001.

[13] Laurens Robinson, 1996. Meta Data: The Key to Managing GIS Data. http://www.esri.com/library/userconf/proc96/TO200/PAP187/P187.HTM

[14] Federal Geographic Data Committee, 1998. Content standard for digital geospatial metadata (revised June 1998). FGDC-STD-001-1998, Washington, D.C.

[15] B. Shneiderman, 1997. Direct Manipulation for Comprehensible, Predictable, and Controllable User Interfaces, *Proceedings of IUI97, 1997 International Conference on Intelligent User Interfaces*, Orlando, FL, January 6-9, 1997, pp. 3339.

[16] Eric Miller, 1998. An Introduction to the Resource Description Framework, *D-lib Magazine*, ISSN 1082-9873, May 1998.
http://www.dlib.org/dlib/may98/miller/05miller.html

[17] Ten Most Common Metadata Errors, FGDC Metadata Education Program, September, 2000. http://www.fgdc.gov/metadata/top10metadataerrors.pdf

[18] DCMI Frequently Asked Questions (FAQ), last access: December 2001. http://dublincore.org/resources/faq/#whatismetadata

[19] An Image Map of the Content Standard for Digital Geospatial Metadata Version 2 - 1998 (FGDC-STD-001 June 1998), United States Geological Survey, Biological Resources Division of USGS. Last access: December 2001. http://biology.usgs.gov/fgdc.metadata/version2/metadata2.htm

[20] Dublin Core Metadata Element Set, Version 1.1: Reference Description, last access: December 2001.
http://dublincore.org/documents/dces/

[21] Kath Stynes, Cartographic Visualization of Population Data, Project Argus, Jan 1996, http://www.mimas.ac.uk/argus/Tutorials/CartoViz/PopViz/

[22] Ana Maria de Carvalho Moura, Maria Luiza Machado Campos and Cássia Maria Barreto, A Metadata Architecture to Represent Electronic Documents on the Web. *3rd IEEE Metadata Conference NIH Campus*, Bethesda, Maryland, U.S.A, April 1999.

[23] C. Lagoze, C. A. Lynch, R. Daniel Jr., The Warwick Framework – A Container Architecture for Aggregating Sets of Metadata, *D-Lib Magazine*, ISSN 1082-9873, July/August 1996.

[24] Ana Maria de C. Moura, Maria Luiza Machado Campos, Cassia Maria Barreto, A Survey on metadata for describing and retrieving Internet resources. *World Wide Web 1*, vol. 4, February 1999.

[25] Susanne Boll, Wolfgang Klas, and Amit Sheth, Overview on Using Metadata to Manage Multimedia Data in *W. Klas, A. Sheth (eds.) Managing Multimedia Data: Using Metadata to Integrate and Apply Digital Data*. McGraw Hill, 1998.

[26] Clifford A. Lynch, The Z39.50 Information Retrieval Standard (Part I: A Strategic View of Its Past, Present and Future). *D-lib Magazine*, ISSN 1082-9873, April 1997.

[27] Pamela Drew and Jerry Ying, A Metadata Architecture for Multi-System Interoperation, *First IEEE Metadata Conference*, Silver Spring, Maryland, April 16-18, 1996. http://www.computer.org/conferences/meta96/drew/metaarch.html

[28] T. Berners-Lee, R. Fielding and L. Masinter, Uniform Resource Identifiers (URI): Generic Syntax, Network Working Group, The Internet Society. August 1998. http://www.ietf.org/rfc/rfc2396.txt

[29] Aidong Zhang and Lei Zhu. Metadata Generation and Retrieval of Geographic Imagery. In *Proceedings of National Conference for Digital Government Research (dg.o2001)*, Los Angeles, California, USA, May 21-23 2001. http://vangogh.cse.buffalo.edu:8080/PAPERS/ZHU2001C.pdf