# AN INVESTIGATION INTO METADATA FOR LONG-LIVED GEOSPATIAL DATA FORMATS

Prepared for the National Geospatial Digital Archive project and funded by the National Digital Information and Infrastructure Preservation Program for Digital Library Systems and Services, Stanford University Libraries by Nancy Hoebelheinrich, nhoebel@stanford.edu and John Banning, jwbanning @gmail.com

Creation Date: 11 March 2008
Adapted for Publication 2 July 2008

Version: 1.1

Status: Final

## EXECUTIVE SUMMARY

As more and more digital data is created, used and re-used, it is becoming increasingly clear that some digital data, including geospatial data created for a myriad of scientific and general purposes, may need to be kept for the long term. What kind of metadata is needed for long term preservation of digital information? Some progress has been made in understanding what policies, treatment, context and explicitly added metadata are important for digital data collections coming from the cultural heritage arena, such as photographic images, encoded texts, audio and video files, and even web sites and the data sometimes derived from interaction with them. Does the experience with cultural heritage digital resources answer the same question for geospatial data?

As a part of the efforts to create the National Geospatial Digital Archive (NGDA), a National Digital Information Infrastructure and Preservation Program (NDIIPP) project funded by the Library of Congress, this paper addresses the question of what kind of information is necessary for archiving geospatial data, and to document research done to answer that question.

This research aims to understand how to best describe those data elements necessary for archiving complex geospatial data as well as what if any, auxiliary data sources are needed for correctly understanding the data. Recommendations for data elements and attributes will be evaluated according to both their logical and logistical feasibility. Building on research done previously within the science dataset and GIS preservation communities, we will suggest necessary metadata elements for the following categories: environment/computing platform, semantic underpinnings, domain specific terminology, provenance, data quality, and appropriate use. Included in the research and analysis will be a comparison of the conceptual models and/or data elements from three different approaches, the content standard endorsed by the Federal Geographic Data Committee (FGDC), the work of the OCLC/RLG sponsored PREMIS work http://www.oclc.org/research/projects/pmwg/ and that of CIESIN, the guidelines for Geospatial Electronic Records (GER). In addition, there will be a discussion of the kinds of information that should be included in a format registry for geospatial materials using a common different geospatial format as an example.

The conclusion drawn from the research is that given both the ubiquity and the comprehensiveness of the FGDC content standard, at this time it is sensible to include the FGDC metadata as part of the submission package along with a PREMIS metadata record (version 3.2), at least for the geospatial formats investigated herein, (ESRI shapefiles, DOQQ's, DRG's and Landsat 7 datasets). The combination of the FGDC metadata and PREMIS goes a long way to satisfy the multiple preservation concepts discussed within the paper, although more research needs to be done with other geospatial and other science data sets to explore how best to use existing elements within the PREMIS Object entity for documenting contextual and provenance information for science data sets.

## Background

As more and more digital data is created, used and re-used, it is becoming increasingly clear that some digital data, including geospatial data created for a myriad of scientific and general purposes, may need to be kept for the long term. As noted in a report from the UK's Digital Preservation Coalition (DPC),

> "The continuing pace of development in digital technologies opens up many exciting new opportunities in both our leisure time and professional lives. Business records, photographs, communications and research data are now all created and stored digitally. However, in many cases little thought has been given to how these computer files will be accessed in the future, even within the next decade or so. Even if the files themselves survive over time, the hardware and the software to make sense of them may not. As a result, 'digital preservation' is required to ensure ongoing, meaningful access to digital information as long as it is required and for whatever legitimate purpose." [1]

For some time, many cultural heritage institutions such as libraries, archives and museums have seen it as their mission to collect, protect and maintain digital collections just as they have done for print-based or "physical" collections. Only recently have other institutions such as the United States National Science Board noted that it is becoming critical to take steps to ensure that "long-lived digital data collections" are accessible far into the future.

In the September 2005 report, "Long-Lived Digital Data Collections: Enabling research and education in the 21st century", the National Science Board's Long-lived Data Collections Task Force undertook an analysis of the policy issues relevant to long-lived digital data collections, particularly scientific data collections that are often the result of research supported by the National Science Foundation and other governmental agencies. From this analysis, the Task Force issued recommendations that the NSF and the National Science Board (NSB) were asked to better ensure that digital data, and digital data collections are preserved for the long-term[2].

Why is it so difficult to preserve digital data? One key factor has to do with the storage of the digital information, i.e., ensuring that the physical bits last over time. The DPC report notes a number of factors that make long term storage of digital information difficult [3] including:

- Storage medium deterioration
- Storage medium obsolescence

---

[1] Waller, Martin and Sharpe, Robert, "Mind the Gap: Assessing digital preservation needs in the UK", published by The Digital Preservation Coalition, York Science Park, Heslington,YORK YO10 5DG, 2006, http:// www.dpconline.org, p. 6.

[2] National Science Board, "Long-lived Digital Data Collections: Enabling research and education in the 21st century", National Science Foundation, September 2005.

[3] Waller, Martin, p. 8.

- Obsolescence of the software used to view or analyze the data
- Obsolescence of the hardware required to run the software
- Failure to document the format adequately
- Long-term management of the data

Storage of the physical bits is not enough as noted by the OCLC/RLG Working Group on Preservation Metadata in a white paper published in January, 2001. As the report states:

> "This, [storage of the physical bits] however, is only part of the preservation process. Digital objects are not immutable: therefore, the change history of the object must be maintained over time to ensure its authenticity and integrity. Access technologies for digital objects often become obsolete: therefore, it may be necessary to encapsulate with the object information about the relevant hardware environment, operating system, and rendering software. All of this information, as well as other forms of description and documentation, can be captured in the metadata associated with a digital object." [4]

The NSF report takes a slightly broader stance, stating that "To make data usable, it is necessary to preserve adequate documentation relating to the content, structure, context, and source (e.g., experimental parameters and environmental conditions) of the data collection – collectively called "*metadata.*[5]" But, what kind of metadata is needed for long term preservation of digital information?

Some progress has been made in understanding what policies, treatment, context and explicitly added metadata are important for digital data collections coming from the cultural heritage arena, such as photographic images, encoded texts, audio and video files, and even web sites and the data sometimes derived from interaction with them. As noted by the DPC report previously cited, knowledge of the format of the digital object is very important. Before data is preserved or archived it is first necessary to understand the formats and/or data types of the information. Comprehension of the format and/or data type of a resource may support re-creation or "re-hydration"of the data at a later date. Such an understanding may also increase the variety of appropriate future uses of the data. Work being conducted by the Global Digital Format Registry (GDFR) aims at capturing this type of information for existing digital formats because current registries do "not capture format-specific information at an appropriate level of granularity, or in sufficient level of detail, for many digital repository activities".[6] Various efforts to create format registries like that of GDFR aim to capture this information, but the scope of these

---

[4] "Preservation Metadata for Digital Objects: A Review of the State of the Art. A White Paper by the OCLC/RLG Working Group on Preservation Metadata", January 31, 2001, p. 4.
[5] NSF Report, p. 20.
[6] "A Registry for Digital Format Representation Information." Stephen L. Abrams and Mackenzie Smith, *DLF Spring Forum*, New York, May 14-16, 2003

efforts typically have not addressed how the elements included in the format registries should be adapted for complex data types such as geospatial.

In the past few years, a number of institutions and organizations have investigated this question. Of special significance recently is the work done by the PREservation Metadata: Implementation Strategies Working Group (PREMIS), another jointly sponsored OCLC/RLG working group. A Final Report and Data Dictionary published in May 2005, "defines and describes an implementable set of core preservation metadata with broad applicability to digital preservation repositories". [7] The PREMIS Data Dictionary (Version 1.0) provides examples of encoded preservation metadata for a number of digital objects, such as a single text document, a slightly more complex object such as an image file and an audio file, and a container file with a file contained within it that also has an embedded file. These examples, and the Data Dictionary are very helpful, but it is not clear that the recommended data elements and data object model will document what is necessary to archive and keep accessible digital data collections of complex data types such as geospatial data, data sets, and databases.

Prior to the work of the PREMIS Working Group, Duerr, Parsons, et al described a comprehensive list of challenges related to long-term stewardship of data, particularly science data. Long-term data stewardship was recognized as having a data preservation aspect but also a requirement to provide both "simple" access and access that facilitated the data's unanticipated future uses. The need for extensive documentation about the data that could support its future uses was noted by Duerr, but also explained in greater detail by several of the references within the article. Specific metadata standards that could be used for documentation were mentioned including the Federal Geography Data Community's content standard and the OAIS Reference model upon which the PREMIS work is closely based. [8]

## Preservation Information for Archiving Geospatial Data

As part of the efforts to create the National Geospatial Digital Archive (NGDA), a National Digital Information Infrastructure and Preservation Program (NDIIPP) project funded by the Library of Congress, the NGDA team has asked what kind of information is necessary for archiving geospatial data. It is the intent of this paper to document the research done in attempting to answer that question.

This research aims to understand how to best describe those data elements necessary for archiving complex geospatial data as well as what if any, auxiliary data sources are needed for correctly understanding the data. Recommendations for data elements and attributes have been evaluated according to both their logical and logistical feasibility. Building on research done previously within the science dataset and GIS preservation communities, we analyze metadata elements for the following categories:

---

[7] "Data Dictionary for Preservation Metadata" from the Final Report of the PREMIS Working Group, May 2005. http://www.oclc.org/research/projects/pmwg/premis-final.pdf . PDF pg. vii.

[8] Duerr R., Parsons, M.A., Marquis, M., Dichtl, R. & Mullins, T. (2004) Challenges in long-term data stewardship. Proc. 21st IEEE Conference on Mass Storage Systems and Technologies. NASA/CP-2004-212750 (pp.47-670). College Park, MD, USA

environment/computing platform, semantic underpinnings, domain specific terminology, provenance, data quality, and appropriate use. Included in the research and analysis is a comparison of the conceptual models and/or data elements from three different approaches, the content standard endorsed by the Federal Geographic Data Committee (FGDC), the PREMIS work, and that of CIESIN, the guidelines for Geospatial Electronic Records (GER). In addition, there is a brief discussion of the kinds of information that should be included in a format registry for geospatial materials using a common different geospatial format as an example.

**Conclusion: From the research and analysis done, we posit that** the existing conceptual approach and data dictionary that the PREMIS group has compiled can be used to describe some complex geospatial data types as long as domain-specific elements from content standards such as the FGDC that extend the PREMIS data elements for geospatial data are used in conjunction.

## Methodology:

### What data is being investigated and why?

For the purpose of this research, four data types were investigated: an Environmental Systems and Research Institute (ESRI) Shapefile, a Digital Ortho Quarter Quad (DOQQ), a Digital Raster Graphics (DRG) image, and a Landsat 7 satellite image. Files of these types are ubiquitous throughout GIS communities and are also readily available for download from the California Spatial Information Library (CaSIL) as well as other GIS clearinghouses. Various complexity levels and different data file types (raster and vector) are reflected in this selection.

### Investigations into various preservation models

As the research and analysis was initiated, the elements contained within the following metadata content standards were compared for their use in geospatial format preservation: the FGDC Content Standard for Digital Geospatial Metadata (FGDC CSDGM) and two preservation data models, the Data Model for Managing Geospatial Electronic Records (GER) and the PREservation Metadata: Implementation Strategies (PREMIS). While the GER data model and FGDC content standard were both developed to focus on geospatial data, PREMIS is designed to be applicable to all archived digital objects. The geospatial specific models, FGDC and GER, differ in their primary objectives. The FGDC is primarily used to aid in the discovery and description of resources or to help identify datasets that may be of use, while the GER "identifies and describes the tables and the fields for storing metadata and related information to improve the electronic record-keeping capabilities of systems that support the management and preservation"[9]. The different purposes of the above mentioned models will be considered throughout this investigation.

[9] Data Model for Managing and Preserving Geospatial Electronic Records Version 1.00 Prepared by: Center for International Earth Science Information Network (CIESIN) Columbia University. June 2005 (http://www.ciesin.org/ger/DataModelV1_20050620.pdf)

The three approaches were compared to discover gaps and overlaps in the following specific preservation concepts or themes: environment/computing platform, semantic underpinnings, domain-specific terminology, provenance, data quality, and appropriate use. Initial investigation into Geography Markup Language (GML) determined that efforts to use GML for archiving geospatial data were in their infancy and too premature to include in this research.

The following section provides an introduction to the models and content standard as well as a visualization of the gaps and overlaps in the data elements. This is followed by a discussion of strengths and weaknesses of each of the investigated models.

**FGDC Content Standard for Digital Geospatial Metadata (CSDGM)**
Rather than a data model, the CSDGM establishes a "common set of terminology for the documentation of digital geospatial data". The standard was developed from the perspective of "defining the information required by a prospective user to determine the availability of a set of geospatial data, to determine the fitness the set of geospatial data for an intended use, to determine the means of accessing the set of geospatial data, and to successfully transfer the set of geospatial data".[10] As stated in Executive Order 12906, 1994, all United States federal agencies using and collecting geospatial data, as well as projects funded from federal government monies, are required to collect or create FGDC compliant metadata. Although it has taken some time, the FGDC CSGDM has become the default metadata standard for most GIS data sets (several desktop GIS application automatically create FGDC metadata records). Additional background information on the FGDC Content Standard for Digital Geospatial Metadata is available at the FGDC website (http://www.fgdc.gov/metadata/meta_stand.html).

**Data Model for Managing and Preserving Geospatial Electronic Records (GER)**
As part of a grant to investigate the management and preservation of geospatial electronic records, the Center for International Earth Science Information Network (CIESIN) has developed a data model, along with cross walks to other standards; an entity-relationship (ER) diagram; and a data dictionary to describe the metadata necessary for the long term retention and management of geospatial data. Included in the grant's work are "appropriate policies, techniques, standards and practices to manage geospatial electronic records". More information on the data model is available in the PDF document prepared by CIESIN (http://ciesin.columbia.edu/ger/DataModelV1_20050620.pdf) and the Geospatial Electronic Records (GER) portal (http://ciesin.columbia.edu/ger/).

**Preservation Metadata Implementation Strategies (PREMIS)**
The PREMIS report and Data Dictionary builds on the Open Archival Information System (OAIS) reference model (ISO 14721)[11], and a Preservation Metadata Framework

[10] Content Standard for Digital Geospatial Metadata. Prepared by: the Federal Geographic Data Committee. FGDC-STD-001-1998 (http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf)
[11] *Reference Model for an Open Archival Information System (OAIS)* (Washington, DC: Consultative Committee for Space Data Systems, 2002), ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf.

developed by an OCLC / RLG working group[12]. To facilitate the logical organization of the metadata elements, and to illustrate its conceptual approach to data, the PREMIS group identified five types of entities: intellectual entities, objects, events, rights, and agents. Definitions of each entity and the relationships among them are described in Section 1 of the Data Dictionary. Specific metadata elements are categorized as belonging or linking to these entities. Several examples are included in the data dictionary to illustrate how to use the preservation metadata; other examples can be found on the PREMIS website. As mentioned earlier, the intention of the PREMIS group was to define elements that were to be considered "***core preservation metadata".*** PREMIS defined "preservation metadata" as "the information a repository uses to support the digital preservation *process*" (emphasis added) while "core" was defined as "things that most working preservation repositories are likely to need to know in order to support digital preservation."[13] Specifically, the PREMIS working group looked at metadata supporting the functions of "maintaining viability, renderability, understandability, authenticity, and identity in a preservation context"[14]. This PREMIS emphasis means that the data dictionary and elements it defines are more narrowly focused than FGDC and GER.

**Data Element Comparison as Differentiated into Preservation Topic Categories**
When brainstorming the need for this research, the NGDA partners came up with a number of concepts that described the type of background information needed for archiving geospatial data including computer platform/environment, semantics, domain specific terminology, provenance, and others. These concepts provided a means to compare the different preservation models and the content standard to determine the strengths and weaknesses of each for preservation purposes.

The preservation concepts are detailed in the tables below. Within each table, details about the concepts or points are presented followed by the terms used by each preservation models / content standard. The FGDC element names are followed by the numbering convention as detailed in the content standard. The GER elements are prefixed with the table name to ensure uniqueness. Where the table remains blank, no element was located that satisfied the criteria.

**1. Environment[15]/Computing Platform**

| Detailed Concepts | PREMIS element | GER element | FGDC element |
|---|---|---|---|
| In what computing environment | creatingApplication | DataFile_FileType | Native Data |

[12] *A Metadata Framework to Support the Preservation of Digital Objects* (Dubin, Ohio: OCLC Online Computer Library Center, 2002), www.oclc.org/research/projects/pmwg/pm_framework.pdf.
[13] PREMIS Final Report, PDF pg. ix.
[14] Ibid.
[15] Environment is defined as characteristics of the hardware and software environment that allow a digital resource to function properly. The approaches taken by the various metadata standards discussed below address different functions such as rendering, viewing, or using the digital resource. Consequently, the elements used to describe the characteristics of an environment will depend upon the function that the data or metadata creator finds important to facilitate through such documentation. It may be important to document more than one environment for a given resource.

| | | | |
|---|---|---|---|
| was the resource created? | | Relationship_Relation | Set (1.13) |
| What software program(s) were used in creating the resource? | creatingApplication/ creatingApplicationName | DataFile_FileFormat | Native Data Set (1.13) |
| What version(s) of the creating software were used? | creatingApplication/ creatingApplicationVersion | DataFile_FileVersion | Native Data Set (1.13) |
| When was the resource created? | creatingApplication/ dateCreatedByApplication | DataFile_Date Modified Provenance_Creation Date | Native Data Set (1.13) |
| What kind of software is required for the resource to be rendered or used (if any)? | environment/software/ swType | Environment_Environ mentType | Technical Prerequisites (6.6) |
| What is the name of software required to view these data, if any? | environment /software/ swName | Environment_Title | Technical Prerequisites (6.6) |
| What is the version of the software required to view these data? | environment /software/ swVersion | DataFile_FileVersion | |
| Are there additional requirements associated with any of the software required to view, render or use these data? | environment /software/ swOtherInformation | Environment_Descript ion | |
| What other software component(s) are needed to make the data functional, i.e. a java class library? | environment /software/ swDependency | Environment_Docume ntation | Technical Prerequisites (6.6) |
| What type of hardware environment is required for the resource to be rendered or used? | environment /hardware/ hwType | Environment_Environ mentType | Technical Prerequisites (6.6) |
| What is the name of the hardware required to view the data (manufacturer, model, version)? | environment /hardware/ hwName | Environment_Title | Technical Prerequisites (6.6) |
| Are there additional requirements associated with any of the hardware required to view, render or use these data? | environment /hardware/ hwOtherInformation | Environment_Descript ion | |

**Comments**:

**GER:** The GER data model contains elements within the Provenance table that capture information about the process used to create a data set while the DataFile table elements capture information about the software used to create each file of the data set. These DataFile table elements include the element “DataFile_FileFormat”, to describe the “Software program used to create the file such as Microsoft Word 2000 and ,Microsoft Excel 2000”; the element “DataFile_DateModified” to describe the “last date and time when file was written or modified”; the element “DataFile_FileType” to describe the “MIME Media Type for file”; the element “DataFile_FileVersion” to describe the “version of the MIME Media Type”; the element “DataFile_FormatRegistry” to describe the “registry to identify the software program used to create or view the file, e.g., PRONOM”; and the element “DataFile_RegistryEntry” to describe the “entry in the Format Registry for the file format”. The GER data model also focuses on describing the

"implementation environment for a data file". This concept, capturing an environment where the data is used, differs from the environment where the data was created.

**PREMIS:** PREMIS defines the "environment" associated with a resource as "the means by which the user renders and interacts" with the content, and makes that element itself a "container"[16] for subelements which allow environments for different purposes to be described. One of the series of related subelements within environment are those which parse creating application information into multiple elements (creatingApplication, creatingApplicationName, creatingApplicationVersion, dateCreatedByApplication) that capture the characteristics of the software (and hardware, if desired) on which the resource was created. PREMIS recognizes the importance of documenting both the creating application and the environment in which the resource can be used, but only requires at least one hardware and software environment where "playable" data is being described.

Other environments recognized by PREMIS that are important for preservation of the resource are those necessary for "rendering", "editing" or other functional tasks associated with using the resource. These purposes can be documented and described using a subelement series that includes environmentCharacteristic, environmentPurpose, and environmentNote. The environment series also has the means to describe both non-software dependencies such as additional components or files (dependencyName, and dependencyIdentifier with its own subseries), as well as software and hardware dependencies as noted in the table above. All could conceivably be used to describe any functional task associated with the data, and the environment that gave rise to the data or is required to perform that function.

Note that changes to a hardware or software environment that affect the digital resource over time are considered out of scope by PREMIS. Thus, it is doubly important to record as much information as possible about the creating or rendering environment that could support the digital resource's future use,

**FGDC:** The optional FGDC content standard element "Native Data Set" attempts to capture a "description of the data set in the producer's processing environment, including items such as the name of the software (including version), the computer operating system, file name (including host-, path-, and filenames), and the data set size". "Technical prerequisites" is used to describe "any technical capabilities that the consumer must have to use the data set in the form(s) provided by the distributor". Although the FGDC content standard categorizes this element with distribution elements that are format specific, the concept is close to what both the PREMIS and GER are gathering, i.e., characteristics of the computing environment where the data properly functions.

**2. Semantic Underpinnings**

| Detailed Concepts | PREMIS Element | GER Element | FGDC Element |
|---|---|---|---|

Formatted: Left

[16] "Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group", May 2005. http:/www.oclc.org/research/projects/pmwg/premis-final.pdf". PDF pg. 2-39.

| | | | |
|---|---|---|---|
| Meaning or essence of the data | N.A. | Provenance_Description | Abstract (1.2.1)<br>Purpose (1.2.2) |
| Significance of the data. Why does the object need to be preserved? | N.A. | Provenance_ReasonForPreservation | Purpose (1.2.2) |
| Function of the data, purpose | N.A. | Provenance_Functionality¸ Provenance_ReasonForCreation | Purpose (1.2.2) |
| Intended community or audience | N.A. | DesignatedCommunity | N.A. |

The sheer number of elements describing the various aspects of a data object (technical, administrative, descriptive) can be overwhelming. Often the documentation of the data is so engrained in details that the most fundamental questions are lost; such as what is the purpose of the data? Why was it created? What does the data represent? This kind of semantic information aims to capture a data set's purpose, abstract and any terminology associated with describing the data (keywords and thesauri), and is especially important for geospatial data.

An example of this necessity is demonstrated with two similar geospatial files representing a street network of the same metropolitan area. The first dataset is the official street centerline file used for emergency management services to locate addresses. It is mandatory for this dataset to contain detailed information on the address ranges within each particular street segment (i.e. "101 -145 Walnut Ave"). The second dataset is cartographic and used for visualization purposes on a tourist map; thus, accurately portraying the topology, angles and geometry of the road network is more important than containing the exact addresses. Without capturing the context for which the files were created and meant to be used, it would be difficult for the user to understand the purpose of the files, thus risking misinterpretation of the data. As there is no inherent information in either dataset about this context, the semantic information about the reasons for the data's existence as well as its uses would have to be contained in the metadata.

While no controlled vocabulary could accurately represent these values, the GER data model and the FGDC content standard support an open ended text field that allows an unlimited space to record this semantic information. The PREMIS data model does not support capturing the justification of the data production. In fact, very few elements that can be considered descriptive elements exist in the PREMIS data model, for two reasons: "First, descriptive metadata is well served by existing standards"…. Second, descriptive metadata is often domain specific." Thus, the PREMIS Working Group recognized that the geospatial domain, for instance, has its own content standards that should be used by those interested in documenting information that is important "both for discovery of archived resources and for helping decision makers during preservation planning.[17].

It is not hard to create statements of purpose such as those data providers would include with the data sets. Examples include

[17] Ibid., PDF pg. 2-3.

*"The main objective for this file is to serve as a reference for mapping projects in NIPC's Regional Geographic Information System (ReGIS). An effort was made to make the graphics consistent with other GIS databases maintained and used by NIPC. The file was intended to facilitate general planning at a regional scale; particular emphasis was placed on collecting main arterials, U.S. and state highways, and maintain an even distribution of roads for general reference."* (From the Northeastern Illinois Planning Commission Major Roads Centerline file) and "*The purpose of this coverage is to be a part of a time series of maps which show property ownership changes in the lower Dungeness watershed from 1863 to 1992*" (From metadata on the Dungeness River Area Property Ownership, 1863)

Authoring statements to define the meaning, significance, or the essence of the data is both a subjective exercise and one that require an intimate knowledge of the data. Furthermore, it is often the case that those person(s) responsible for data documentation or creation of metadata do not have a thorough understanding of the data. The significance of the data may differ among the data users, authors, and metadata creators. Where the original authors may have had a specific intention for the data, "to be used to delineate tax parcels", for example, scientists may later see additional uses unknown at the time of creation. Some of these uses that future scientists may wish to apply the data may well be inappropriate, resulting in errors and misinformation. Arguably, collecting information about the "designated user community" for a given data set or collection is a very important responsibility for a data archive. [18]

There is no question that the creator's original intention for the data is valuable and should be kept when provided. This semantic information offers not only context but also insights into limitations that may not otherwise be explicit. The FGDC content standard recognizes this importance and has made both the abstract (a brief narrative summary of the data set) and the purpose (a summary of the intentions with which the data set was developed) elements required. These requirements support the primary purpose of the FGDC content standard, i.e., discovery and identification of geospatial resources, but are not a core tenet per se for generic preservation of resources as defined by the PREMIS specification.

**3. Domain Specific Terminology**

| Detailed Concepts | PREMIS element | GER element | FGDC element |
|---|---|---|---|
| Theme Keywords | N.A. | Provenance_Description | Theme Keywords (1.6.1) |
| Spatial Coverage | N.A. | Provenance_SpatialCoverageDesc | Place Keywords (1.6.2) |
| Time Period | N.A. | TemporalData_TemporalStart<br>TemporalData_TemporalEnd<br>TemporalData_TemporalDescription | Temporal Keywords (1.6.4) |
| Stratum | N.A. | N.A. | Stratum Keywords |

[18] "Designating User Communities for Scientific Data: Challenges and Solutions", Mark A. Parsons and Ruth Duerr, National Snow and Ice Data Center/World Data Center for Glaciology, Boulder, Colorado. *Data Science Journal*, Vol. 4 (2005) pp. 31-38.

| | | | |
|---|---|---|---|
| | | | (1.6.3) |

Geographic technical terms are not limited to subject matter terms such as "transportation", "hydrography", or "parcel". Geospatial data are unique in that the data are associated with locations. These locations may be portrayed either through place names ("New York, New Amsterdam"), spatial coordinates (latitude/longitude) and coordinate ranges, or both. In addition to location information, geographic data are often acquired as a snapshot at a certain time. Therefore, in addition to topical keywords, temporal, spatial, as well as stratum keywords are often necessary to accurately portray the data.

The FGDC content standard creators understood that geospatial data represent an abstraction of a place or area at a given time, typically dealing with a theme. The FGDC standard allows for that information to be captured in various metadata elements. Related concepts may be described in a number of ways, and an unlimited number of times in the various keyword concepts (theme, place, stratum, temporal). Citation of a formally recognized thesaurus is also supported to help further understand the terminologies used to describe the data. An example of this methodology is using a specific biological taxonomy for a data set that captures the distribution of the species.

Although not as inclusive as the FGDC content standard, the GER also sees the importance of recording the various vocabularies used to describe geospatial data. The data model supports the following data concepts through database attributes and relational database tables: spatial coverage, thematic keywords, and time period. Although a relational database structure, the GER may be limited in the way theme keywords and the spatial coverage are recorded as it is not clear whether these fields support an unlimited number of entries as would seem necessary.

Because PREMIS is a generic data model for the preservation of all types of resources, it does not accommodate those concepts that are particular to geospatial data. Furthermore, descriptive metadata elements are not included in PREMIS which precludes the inclusion of subject or theme keywords. Instead, PREMIS assumes that such descriptive information would be recorded using a more domain specific metadata schema such as FGDC or GER.

**4. Provenance**

| Detailed Concepts | PREMIS element | GER element | FGDC element |
|---|---|---|---|
| Information about the events, parameters, and source data which constructed the data set prior to archival ingestion, and which need to be retained.. | Object Entity<br>environment<br>significantProperties | Provenance Table<br>Origin<br>Version<br>PreIngest<br>CreationDate<br>DesignatedCommunity<br>ReasonForCreation<br>CustodyHistory | Process Step (2.5.2)<br>Process Description (2.5.2.1)<br>Source Used Citation (2.5.2.2)<br>Process Date (2.5.2.3) |
| Source from which the information was derived? | Object Entity<br>Relationship/ | Provenance Table<br>Origin | Source Information (2.5.1) |

| | | | |
|---|---|---|---|
| | relatedObjectIdentification | ProvReference Table | |
| Changes, modifications to the data inside the preservation archive | Event Entity<br>eventIdentifier<br>eventType<br>eventDateTime<br>eventDetail<br>eventOutcomeInformation<br>linkingAgentIdentifier<br>linkingObjectIdentifier<br>Agent Entity<br>agentIdentifer<br>agentName<br>agentType<br>Object Entity<br>linkingEventIdentifier | Provenance Table<br>Relationship Table<br>ProvenanceNote Table<br>Person Table<br>Institute Table<br>Document Table<br>Property Table<br>Identification Table<br>ProvReference Table | |

Alterations, versions, and the various processes and revisions that went into creating data sets are all considered contextual information worthy of documentation. Not only does this type of detailed history support the re-creation of the objects but it also documents the considerations and thoughts that went into its creation. The suggestion has been made that this kind of information is especially important for science data sets, particularly for supporting unanticipated future uses of the digital resource. One of the requirements for science data sets that is described in the Duerr, Parsons article is the necessity to extensively document characteristics of the creation of the data set such as the identification of instrument / sensors, its calibration and how that was validated, the algorithms and any ancillary data used to produce the resource.[19] In the science community, according to the Duerr, Parsons article, such information is considered "provenance" or processing history.

GIS data sets often require numerous processes, command, and/or tools to create the final product, thus it is important that the elements documenting them are repeatable. Consider the creation of a demographic map as an example. Before such a map is published, numerous datasets may have been combined or merged together, re-projected into the appropriate coordinate system, and then an agreed upon classification system is applied to the result. The decisions and processes that led to the creation of the map are examples of the types of information that are captured within data lineage. For instance in the following phrase:

```
Merge c:\temp\states1;c:\temp \states2; c:\temp\USA
```

not only is the command, or process used to create the output documented, but also the input data sources.

Today, improved GIS technologies (ESRI's ArcGIS) often capture these specific command histories and other details (dates, environment) used to generate the output data set. This information can be captured in numerous ways through custom code, but by

[19] Hunolt, Greg. "Global Change Science Requirements for Long-Term Archiving. Report of the Workshop", Oct 28-30, 1998, USGCRP Program Office. March 1999.

default is written into lineage elements within geospatial metadata standards such as the FGDC content standard. The FGDC attempts to record the decisions, commands, and processes that go into the product throughout the life cycle. Within the content standard, the Data Quality section contains metadata elements specific to providing information about these choices. The data sources and process description elements can be quite long when complete data creation details are provided (Appendix B).

In the GER model, elements in the DataFile table capture information about the software used to create each file, and elements in the Environment table capture information about the implementation environment. In addition, the GER provides several opportunities to record the purpose for creating and the processes used to create data objects prior to accession into an archive. The latter elements are all found within the Provenance table and include; Origin, Version, PreIngest, CreationDate, DesignatedCommunity, ReasonForCreation, and CustodyHistory. The GER model is focused on "the history and changes that occur during the entire lifecycle of an object" and specifically on accession into the archive and changes to the data object after the data has been ingested. It is not focused upon the history and processes that were used in the original data development. As a result, there are significantly fewer elements in this model than are provided by the FGDC content standard when documenting data lineage. Also, it is unclear whether it is permissible to have repeating entries in the GER Provenance table for each data object as is allowed in the FGDC standard.

The PREMIS data model uses a number of entities to record pertinent information about the data object both prior to its ingestion into a digital repository and after the data object has been ingested and preservation actions have been taken on it, such as migration from one format to another. For instance, this kind of information could be included within the environment container element within PREMIS. It remains to be seen how feasible this element and its subelements would be for science or geospatial data sets since the emphasis is upon hardware and software, neither of which would really cover the types of contextual information described above.

Important features to retain can be described within the "significantProperties" element of the Object entity as well within the environment container element within PREMIS Object as described above. One important factor to note is that most of the PREMIS metadata elements can be used to describe data objects at several levels of decomposition including at the representation, file or a bitstream levels. A representation could be considered as an abstract, ideal or intellectual entity composed of files or bitstreams, while a data object could be described at the single file or bitstream level. This data model provides a great deal of flexibility in describing a number of levels or layers of which a data object could be composed including a "relationship" element that allows explicit descriptions between and among layers or levels of a data object. In addition, PREMIS provides for Event, and Agent entities thus enabling a data provider or digital repository staff the means to describe important events and software, organizations, and/or individuals which / who have had a significant role to play in the provenance or lineage of a data object. While the creation of the PREMIS metadata that records this kind of information would not be trivial to include, especially if done manually, it could

be an important means for describing changes and/or modifications to a data object that occurs prior to and after its ingestion into a preservation repository.

**5. Data Trustworthiness**

| Detailed Concepts | PREMIS elements | GER elements | FGDC elements |
|---|---|---|---|
| Who are the parties responsible for the creation, development, storage and/or maintenance of the data set. | Agent Entity<br>agentIdentifer<br>agentName<br>agentType | **Institution Table:**<br>Institution _Institution<br>Institution _InstitutionRole<br>Institution_ InstitutionName<br>**Person Table**<br>Person_PersonRole<br>Person_FirstName<br>Person_LastName | Originator (8.1) |
| Where is the data available? (Location) | Object Entity<br>objectIdentifier<br>storage<br>contentLocation<br>storageMedium | **Distribution Tables:**<br>Distributor Table<br>Dissemination Table<br>DissemAltPoint Table<br>ProvDissemination Table<br>PDFileList Table<br>Catalog Table<br>CatalogEntry Table | Distributor (6.1)<br>Resource Type (6.2)<br>Distribution Liab.(6.3)<br>Ordering Process (6.4)<br>Technical Prereq (6.6) |
| How is the data available? What important factors about the data should be preserved? | Object Entity<br>objectCharacteristics<br>format<br>significantProperties<br>environment<br>dependency | **Distribution Tables:**<br>Distributor Table<br>Dissemination Table<br>DissemAltPoint Table<br>ProvDissemination Table<br>PDFileList Table<br>Catalog Table<br>CatalogEntry Table | Ordering Process (6.4)<br>Technical Prereq. (6.6) |

**Comments**: "For a scientist to be able to trust that the data have not been changed the scientist must be able to trust that the preservation practices of the source of the data are adequate; that archive media are routinely verified and refreshed, that the facilities are secure, that processes to verify and ensure the fixity of the data are operational, that geographically distributed copies of the data are maintained as a protection against catastrophe, and that disaster recovery plans and procedures are in place."[20]

As mentioned in the Duerr, Parsons article, and corroborated by other discussions on "trusted digital repositories"[21]a data set's integrity, or the confidence that the data is accurate and correct, is correlated with trust in those who created the information, as well as those who've stored the data. Data from an unreliable or unknown source is often passed over for the same information from a more trustworthy source. Because of this, recording the party responsible for the creation, adaptation, storage and/or maintenance of the data is considered valuable,

[20] Ibid., p. 113.
[21] See for example, "Trusted Digital Repositories: Attributes and Responsibilities, An RLG-OCLC Report", Research Libraries Group, first published in May 2002, http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf.

In addition to recording information about the parties responsible for the data, the media on which the data is captured or stored can also be helpful (i.e. DVD, CD, network download, tape). Most data seekers would not find it useful to go through the exercise of finding pertinent data only to realize that the media and available players are incompatible. The information about capture media is less important for the data that is being stored in a preservation repository, of course, as presumably, the repository can be presumed to store the data on long-term reliable media. In these cases, trustworthiness of the repository in which the data is located may be more important than the medium upon which it is stored.

Each of the data models and content standards provide the ability to capture such information, but through different means.

FGDC: Capturing information about the data's creator is done by use of the originator element which holds the "name of an organization or individual that developed the data set." To describe the variety of resource types available for a particular data set and how to obtain them, FGDC provides a "Distribution" section which may be repeated to provide for the various available media types. Included within the distribution section is an element used to declare any "technical prerequisites" that may be needed for the execution of the data for a particular resource type.

GER: Similarly to FGDC, the GER provides a set of elements within several distribution tables that can be repeated. The GER also provides the means to capture the data originator or creator. This may be accomplished by creating an entry for the data creation party, and then declaring the relationship between the party and the data in the relationship table.

PREMIS: Generally, PREMIS would regard facts about both the originator of a data set and its media resource types as descriptive information, thus does not provide a specific means for recording the information. Indeed, the PREMIS Working Group did not consider at length events or processes that occur before ingest and was not convinced that these were core knowledge for a preservation repository." [22]Rather, given the emphasis upon preservation, PREMIS provides the means to capture provider, format and location information about the data once it is in a preservation repository using the Agent and the Object entities. PREMIS also provides an element called "significantProperties" that can be used to record characteristics of a particular data set that are important to be preserved, such as objective technical characteristics. See more discussion of this element in the next section on Data Quality.

**6. Data Quality**

| **Detailed Concepts** | **PREMIS elements** | **GER elements** | **FGDC elements** |
|---|---|---|---|
| General condition statement. | Object Entity objectCharacteristics | Provenance_RecordCondition | Completeness Report (2.3) |

[22] Ibid, PDF p 4-12.

| | | | |
|---|---|---|---|
| | significantProperties | | |
| Accuracy of the identification of entities and assignment of attribute values in the data set" | N.A. | N.A. | Attribute Accuracy (2.1) |
| Explanation of the fidelity of relationships in the data set and tests used" | N.A. | N.A. | Logical Consistency Report (2.2) |
| Assessment of the accuracy of the measurements taken, e.g., where the center point of each pixel is located. | N.A. | N.A. | Positional Accuracy (2.4) |
| Description of how far apart individual measurements were taken, e.g., the size of each pixel. | N.A. | Provenance_Spatialresolution | N.A. |

**Comments**: The quality of the data often determines its usefulness for a particular purpose, e.g., is this coastline dataset detailed enough for navigation? The quality of spatial data sets is often a function of spatial resolution, i.e., how far apart individual measurements are, positional accuracy, i.e., how accurately each position is known, or measurement accuracy. While the FGDC content standard provides numerous data condition elements, ranging from an attribute's accuracy to cloud coverage percentage, the other data models provide generic catchall elements for data quality.

Within a preservation context, determining data quality often falls within the curatorial assessment function of a preservation work plan. This work is usually completed before the decision about whether data is to be ingested into an archive and preservation activities (such as format assessment, metadata development and collection) have begun. The act of deciding that the data is in a condition worthy of preservation, and that the collection is significant enough to retain infers that some standards for minimum data quality or importance have been met, although explicit inclusion of appraisals or other selection / evaluation tools completed by the collecting institution would be valuable to include with the data.

For science data sets, this may well mean that contextual information about the creation of the data set such as instrument calibration or research questions being hypothesized and addressed, meanings of column and row headings in a tables within the data set, etc. need to be evaluated and ideally, included with the data set prior to selection for inclusion into a preservation repository, and arguably, for proper use of the data. As discussed in the "Environment" section above, the existence and completeness of this kind of information is very important to include for data sets. This kind of information *could* be considered descriptive metadata, thus explaining the generic approach taken by both GER and PREMIS. For geospatial and GIS data, however, it is important documentation to accompany the data set within the preservation repository, critical to preventing data misuse.

FGDC: Besides the data quality elements noted in the chart above, FGDC also provides other elements that extend the typical data quality statements by providing specific metadata elements relating to GIS data sets. These include attribute accuracy (mandatory if applicable), completeness (mandatory), and data lineage (mandatory). The FGDC

also supports raster GIS data sets that may have clouds obscuring the imagery by supplying an element to capture the percentage of cloud coverage.

GER: To record the specifics related to the decision of retention and preservation, the GER data model contains repeating elements within the Property table. These elements are intended to capture quality review information. They include the “PropertyName” element to record the “Name of the property describing an object”, the “PropertyDesc” element to record the “Description of a property”, and the “PropertyStatus” element to record the “Current status of the property”. Suggested values for the “PropertyName” element include “Quality Review Pending”, “Quality Review Complete”, and “Failed Content Quality Review” to facilitate the quality review process. In addition, the GER data model contains a Decision table of elements to describe each "decision that affects the provenance or dissemination for one or more objects" and the ProvenanceDecision table containing elements to describe each data set "affected by a particular decision". The GER data model also contains the ProvReference table to record information about publications, such as peer-reviewed articles, that refer to a data set.

The GER data model does support two metadata elements that capture the “condition of record” and the spatial resolution. Although these elements map back to similar elements found in the FGDC data quality section, the GER data model does not provide the same level of detail as does the FGDC model in determining quality. However, the GER data model provides capabilities in its Document table to describe, capture, and manage the content of various documents that describe a data set, including documents containing FGDC compliant metadata, user guides, and documents that conform to other standards. If part of the purpose for providing metadata is to better equip users to make informed decisions about using the date, given the subjectivity inherent in such judgments, more opportunities for documenting various properties of the data quality are welcome.

PREMIS: While no metadata elements in PREMIS specifically address the quality of geospatial data, there is a means to record subjective judgments about characteristics of data that should be preserved over time. The significantProperties element within the objectCharacteristics container is included in the data model in order to address technical properties of a file or bitstream that should be preserved for future presentation or use. It is possible to apply the signficantProperties element to different aspects or layers of a data set, or to the entire resource. For instance, it may be very important to the use of a data set for a specific purpose that a JavaScript included with the data set be retained for purposes of rendering it. With this requirement documented in the significantProperties for either the set or a specific file or bitstream component of the data set, it would be easier to document any activities or “Events” that occur during migration of the data set to a different format should that be necessary or desired. Probably the best use of this PREMIS element would be in conjunction with the more geo-specific and explicit elements provided by FGDC or GER.

**7. Appropriate Use**

| **Detailed Concepts** | **PREMIS elements** | **GER elements** | **FGDC elements** |
|---|---|---|---|
| Legal use and liability statements | Rights entity | | Use Constraints (1.8) |

| | | | |
|---|---|---|---|
| | permissionStatement<br>Agent entity | | |
| Technical characteristics related to data type / format that impact use | Object entity<br>objectCharacteristics / format<br>formatRegistry | DataFile_FormatRegistry<br>DataFile_RegistryEntry | |

**Comments**: Two aspects of the appropriate use of data are important for this discussion. First is the capability of including with the resource an explanation or reference to the legal terms associated with its use. Two of the data models, PREMIS and FGDC, provide the means to record this information. The other aspect of appropriate use has to do with the technical characteristics of the data that make it simply more effective or accurate for one or more uses than for others. Both usage aspects could apply either to an instance of a specific data type/format (e.g. a municipality's pipeline shapefile compared to a shapefile of the world coastlines) or to the entire data type/format itself (e.g. Landsat7).

**Legal Use / Liability Statements:**

FGDC: The FGDC defines the use constraints element as "restrictions and legal prerequisites for using the data set after access is granted". These include any usage constraints applied to assure the protection of privacy or intellectual property, and any special restrictions or limitations on using the data set." Often the use constraint element contains a liability statement protecting the data provider from lawsuits due to possible inappropriate uses. The following represents a typical liability statement found in the FGDC use constraint element for a specific instance of a data format:

> *Planimetric maps should be used for intended purpose and should not take the place of ground surveys for highly detailed requirements. The information and depictions herein are for informational purposes only and Waukesha County specifically disclaims accuracy in this reproduction and specifically admonishes and advises that if specific and precise accuracy is required, the same should be determined by procurement of certified maps, surveys, plats, Flood Insurance Studies, or other official means. Waukesha County will not be responsible for any damages which result from third party use of the information and depictions herein, or for use which ignores this warning.* (**http://www.waukeshacounty.gov**)

PREMIS: The PREMIS data model provides the capability of recording or associating statements about rights and permissions related to a resource via the Rights entity. PREMIS defines "rights" and "permissions" [23] more broadly than FGDC's focus upon "use constraints", but only defines "minimum core" rights and permissions to be those granted to a repository necessary to perform its preservation function. There is no restriction on using the Rights entity to record other kinds of rights and permissions such as any constraints on use. Using the "permissionStatement" container, it is possible to

[23] "Rights are entitlements allowed to agents by copyright or other intellectual property law. Permissions are powers or privileges granted by agreement between a rightsholder and another party or parties." Ibid, PDF, p. 2 -88.

include specific information about the permissions granted, any restrictions upon the permission, and/or links to a granting Agreement which fully documents the rights and permissions, uses and constraints upon the resource, in a manner similar to what is possible with FGDC,. It is also possible to use the Agent entity in conjunction with the Rights entity to identify those who can grant permissions and rights, if desired.

**Technical Characteristics of Format / Datatype:**

In terms of the technical aspects of a data format that govern its appropriate use, some may argue that resolution and spatial accuracy of a discrete resource are the most significant drivers of the appropriate use of geospatial data, but there are other factors that should be considered such as time period of content and attribute accuracy. Both GER and FGDC content standard provide places for such values in other elements previously described. As well, FGDC allows overall comments on the use of the data.

In addition however, descriptions of the appropriate use of large ubiquitous data products (DRGs, DOQs, and Landsat7 imagery) should be managed at an authoritative location, such as a format registry, designated government agency or national data center. Awareness of the suitable uses of the data parallels the need to be familiar with the different sensor specifications and satellite configurations which also could be managed in the format registry. See a brief investigation of technical characteristics of an ESRI shapefile in Appendix C.

As delineated above, both PREMIS and GER provide elements to describe or link to entries in format registries when those are available.

The need to obtain and document the data's appropriate usage is just as important as the environmental characteristics that make the data "play-able". Whether that understanding is explicitly stated in each data instance, such as an FGDC metadata record, or contained within a registry for an entire data product is dependent on similarities of the data type/format. While commonalities among DRGs, DOQs, and Landsat7 data sets may support the use of one use statement for an entire data collection or data set, shapefiles should be evaluated on a case-by-case basis due to their variability.

## Discussion of strengths / weaknesses

**FGDC Content Standard Strengths / Weaknesses**

The most obvious strength of the FGDC content standard is its richness and specificity. The standard contains a mixture of what is traditionally considered descriptive and technical metadata that is designed specifically for geospatial and GIS materials. As such, it is a very important contribution to the comprehensive metadata of a "geo-resource." In addition, a large user community has adopted the FGDC metadata standard, aided by the requirement that FGDC metadata accompany the data resources provided by all United States federal agencies as well as scientists and organizations funded by the U.S. government. As a result, a significant number of data sets today are accompanied by FGDC metadata, although often with only the minimum number of elements

populated for each document (e.g., only the abstract, purpose, and key descriptive elements).

The richness of the FGDC metadata content standard could also be considered a weakness as the number of metadata elements can be overwhelming and confusing to use. The complexity of the standard and a resistance to metadata creation in general combine to result in the tendency for FGDC records to contain only the minimum number of elements completed. These minimalist FGDC records *may* be sufficient for data discovery and description, but may well be less than satisfactory for long term preservation, especially if complex or compound resources are being described. Those elements in FGDC that document the context of a data resource and the specific applicability for given uses intended by the data creator(s) would be of special emphasis for long term preservation, as previously discussed.

Two other areas of explicit documentation could be considered of particular importance for long-term preservation of data resources, and neither of these presently are part of the FGDC content standard. The first has to do with the ability to explicitly describe structural relationships among the components of a data resource, of great importance for complex or compound data resources. Some capability exists within FGDC for describing relationships among the metadata records of objects, but only in terms of 'single inheritance', i.e., a parent to child. The second weakness of FGDC from a preservation point of view is an emphasis upon recording the state of the resource at the time of creation with little opportunity to describe important events in the lifecycle of the resource as it is managed and preserved over time.

Even though FGDC metadata records are often incomplete, it is true more often than not that the record contains some information of value for preservation. Many of the elements detailed above are required (purpose, abstract, theme keywords, access constraints), for instance, and must be present. Other optional FGDC metadata elements contained in the standard strengthen and aid in preservation practices.

Like many metadata standards developed for broad based user communities, customization of the standard has occurred to fit the needs and policies of given user communities. An example of a customization of the standard can be seen in the development of metadata profiles such as that of the ESRI profile of the FGDC Content Standard for Digital Geospatial Metadata. While the "objective of this profile is to make metadata more accessible and useful on a daily basis when browsing, searching, and managing data"[24], several additional elements seen as valuable for preservation purposes were included. These include elements such as dataset size, language of the data set, native dataset format (i.e. dBASE Table, Shapefile, Text File), attribute type, attribute width, attribute indexed, and process software and version (used in documenting the data lineage). Some of the most meaningful metadata elements the ESRI profile provides are those relating to raster images; cell size direction, cell size units, bits per pixel, compression type, image color map, and raster origin. While these elements may be

[24] "ESRI Profile of the *Content Standard for Digital Geospatial Metadata"* Copyright © 2001–2003 ESRI. p.4 (http://downloads.esri.com/support/whitepapers/ao_/GeospatialMetadataProfile_J8709_3-03.pdf)

considered technical rather than core preservation elements, they are necessary in documenting the environments which the data was created and utilized. The ESRI profile was designed to align with International Organization for Standardization (ISO) 19115, *Geographic Information—Metadata.*

The FGDC standard was the foundation on which the ISO 19115 metadata standard was built. The ESRI Profile is in part intended to facilitate the creation of ISO metadata by including some elements that were proposed for the ISO standard for which information could be automatically harvested from spatial datasets. When the US National Profile of the ISO standard is adopted to replace the FGDC, ESRI will design a profile of the ISO standard so that properties of datasets can continue to be harvested and recorded in metadata documents. Further documentation on the ESRI profile of the Content Standard for digital geospatial metadata can be found at the following website: http://www.esri.com/metadata/esriprof80.html.

**GER Strengths / Weaknesses**

The GER work done by the CIESIN is focused on the preservation and long term management of digital geospatial objects. The GER effort was intended to delineate a structure for managing geospatial digital resources in a relational database, thus providing a means for implementation. The entity relationship (ER) diagram accompanying the data model contains thirty-nine tables classified into five categories that closely adhere to preservation concepts CIESIN determined that the digital preservation community has adopted: organization, provenance and attributes, administration, distribution, and physical properties. The analysis was done by comparing various preservation metadata standards and getting input from members of numerous advisory boards, and is quite comprehensive.

The only area of weakness that seems evident with the GER model is the difficulty in using the elements of the model at anything but the physical level of the file or files that comprise a geospatial resource, e.g., not at an abstract or intellectual level. In addition, while the GER model does include means of describing relationships among physical components of a resource by means of a Relationship table, it might be difficult to describe relationships that are not hierarchical in nature due to the GER relational implementation structure. It would be useful to see and understand how one would describe relationships among a complex resource using the GER model.

As thorough as the GER model is, it is still in its infancy in terms of its use within the geospatial community. To date, there has been no implementation; thus, it is unknown how well the model will work with the myriad of GIS datasets that it was created to support. An established user community has yet to develop, and no best practices documentation for preserving different data types is available. An example of this is the absence of the conditionality (required, mandatory if applicable, optional) of different fields in the database. Because there is a lack of practical implementation, it is entirely up to each individual implementer to decide whether fields need to be populated.
GER offers various crosswalks to other metadata standards used for preservation, as shown below. Since the GER data model was developed to be complementary to the

FGDC Content Standard for Digital Geospatial Metadata (CSDGM) as well as to other standards that describe discovery or descriptive metadata, a crosswalk has not been created between the GER data model and the FGDC CSDGM.

**Metadata standards included in the GER crosswalks**[25]

| | |
|---|---|
| *e-Government Metadata Standard Version 3.0*, | http://www.govtalk.gov.uk/schemasstandards/metadata.asp |
| *Model Requirements for the Management of Electronic Records: MoReq*. | http://www.govtalk.gov.uk/schemasstandards/metadata.asp |
| *DOD 5015.2-STD Design Criteria Standard for Electronic Records Management Software Applications*. | http://www.cornwell.co.uk/moreq.html |
| Dublin Core Metadata Initiative. *Dublin Core Metadata Initiative Metadata Terms*. Adopted as Information and documentation – The Dublin Core metadata element set (ISO 15836:2003) and as The Dublin Core Metadata Element Set | http://jitc.fhu.disa.mil/recmgt/standards.html |
| National Library of Australia. *Preservation Metadata for Digital Collections: Exposure Draft*. | http://www.dublincore.org/ |
| National Library of New Zealand. *Metadata Standards Framework, Preservation Metadata*. | http://www.natlib.govt.nz/files/4initiatives_metaschema_revised.pdf |
| Online Computer Library Center (OCLC) and Research Libraries Group (RLG). *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group*. | http://www.oclc.org/research/projects/pmwg/premis-final.pdf |

The GER data model is an attempt to describe the information necessary to manage a digital object repository by creating a schema for data management of objects throughout their life cycle. Less emphasis is placed on understanding the necessary metadata elements that are specific or unique to the preservation of geospatial data. As a result, it is sometimes unclear why some metadata elements have been included while others that other preservation schemas have included have been ignored.

While the relative number of users of the GER data model is unknown, the model is rather flexible and has promise for being an important research and analysis tool in understanding geospatial archives. The developers of the model encourage "adoption of the data model or a subset of the tables and fields" that may be "improved to foster management and preservation of digital objects and collections." As a user community develops, an understanding of the weaknesses and strengths of the GER data model will emerge and undoubtedly be reflected in later versions of the model.

[25] Data Model for Managing and Preserving Geospatial Electronic Records Version 1.00. Prepared by: Center for International Earth Science Information Network (CIESIN). Columbia University. June 2005 (http://www.ciesin.org/ger/DataModelV1_20050620.pdf)

**PREMIS Strengths / Weaknesses**

The PREMIS data model is designed to apply to *all* archived digital resources. The PREMIS Working Group conducted extensive comparisons with other efforts to define preservation metadata, and ultimately decided to focus upon delineating and defining only those elements considered "*core*" for the preserving of digital resources at various stages of its lifecycle. As a result, descriptive metadata, which is arguably necessary to completely understand an object, is largely excluded in PREMIS. As previously discussed, this includes the semantic information that captures a data set's purpose, an abstract and any of the terminology that is especially important for geospatial data. From the point of view of full preservation of geospatial data, this is a weakness of the PREMIS element set.

There are important strengths inherent in PREMIS, however, that make it an important contribution to full and long term preservation of geospatial data. First, the fact that PREMIS can be applied to both abstract and actual or "physical" components of a resource is a valuable concept. This approach allows preservation metadata to be collected using a flexible method of attribution for both the intellectual and the physical aspects of a digital resource. At the very least, with PREMIS it is possible to document the different files that are used in creating complex objects as well as their relationships. Many geospatial datasets are composed of various files that must interact to render the correct geospatial abstraction (i.e. shapefiles, DOQQ, DRG, Landsat). The PREMIS model provides the ability to document these relationships through use of the "relationship" semantic unit.

The second important strength of PREMIS is its capability for describing actions taken during the lifecycle of the resource. Thus, as a resource ages and is migrated to different media, formats, or archives for continued use over time, it is possible to record changes in the important characteristics of the object, events that have happened, and who or what agent had a role in these events. This kind of information could be critical for continued rendering, use, or other functions associated with the resource.

**Conclusion and Recommendations for Geospatial Metadata Preservation**

While there are sufficient means in both the GER and PREMIS for capturing most of the preservation concepts, additional elements are needed to fully document a dataset's context. Without disclosure of the purpose for the data or what it represents, a lack of confidence or uncertainty will remain; therefore, it is recommended that metadata about the semantic underpinnings and data quality, when available, accompany geospatial datasets for preservation.

Both the ubiquity and the comprehensiveness of the FGDC content standard, including the mandatory requirement of key descriptive metadata elements (abstract, purpose, and keywords) that provide semantic context make it sensible to include the FGDC metadata as part of the submission package along with a PREMIS metadata record, at least for the geospatial formats investigated herein, (ESRI shapefiles, DOQQ's, DRG's and Landsat 7 datasets). The combination of the FGDC metadata and PREMIS significantly satisfies the

multiple preservation concepts previously discussed (environment / computing platform, semantic underpinnings, domain specific terminology, provenance, data quality, and appropriate use).

The prevalence and availability of FGDC metadata is a factor in this recommendation. As previously noted, many commercial GIS software packages by default provide initial spatial metadata as well as tools that make data documentation easy. Most GIS professional have at least an initial exposure to geospatial metadata concepts and terminology and understand correct data documentation. When a complete FGDC metadata record is available for a digital resource, it would seem that very little is needed in terms of preservation metadata that the PREMIS data models offers, unless one wanted to describe the resource at different levels, such as the more abstract level of the resource as an intellectual entity (representation), or at the component level such as a file and/or bitstream. In these cases, PREMIS provides the means for explicitly describing the relationships among these levels by means of the "relationship" element. It may also be possible to use the relationship element to associate related files or websites for example, that provide more of the contextual information important for geospatial and other science data sets. It would be important to test this application of PREMIS with a variety of geospatial and science data sets.

Also importantly, the data structures inherent in both FGDC and PREMIS provide a means for managing objects once they have entered the archive, especially given the Event and Object entities within PREMIS. This combination might, over time, provide the best option for an institution depending upon how the data models and the XML technology upon which they are based fit implementation and preservation strategies. Of course, PREMIS too has not yet been tested over a long time period. Since the purpose of the PREMIS model is to record preservation information generic to all types of digital data, and thus not those qualities specific to geospatial data that GER provides, each digital archive may need to assess whether GER or PREMIS would be more useful for them. This decision depends upon the variety of digital resources being collected in the archive as well as the implementation technology being used, e.g., RDBMS vs. XML.

# Appendix A - Shapefile representation using PREMIS data model

## 1. shp file

```
<?xml version="1.0" encoding="UTF-8"?>
<ROOT>
<objectIdentifier>
        <objectIdentifierType>SDR_</objectIdentifierType>
        <objectIdentifierValue>shp_07108e3d-5fd1-11da-b211-19e7a5cf4814</objectIdentifierValue>
</objectIdentifier>
<preservationLevel/>
<objectCategory>file</objectCategory>
<objectCharacteristics>
        <compositionLevel>0</compositionLevel>
        <fixity>
                <messageDigestAlgorithm>MD5</messageDigestAlgorithm>
                <messageDigest>5a00388f35ac9fc20fe8f11026548f74</messageDigest>
                <messageDigestOriginator>Stanford Digital Repository</messageDigestOriginator>
        </fixity>
        <size>54280</size>
        <format>
                <formatDesignation>
                        <formatName>shapefile</formatName>
                        <formatVersion>1.0</formatVersion>
                </formatDesignation>
                <formatRegistry>
                 <!--NOTE: This is a placeholder as this format registry does not yet exist.-->
                        <formatRegistryName>NGDA Format Registry</formatRegistryName>
                        <formatRegistryKey>http://www.ngda.org/format/def/shapefile/DBase.html</formatRegistryKey>
                        <formatRegistryRole>Specification</formatRegistryRole>
                </formatRegistry>
        </format>
        <significantProperties/>
        <inhibitors>
                <inhibitorType/>
                <inhibitorTarget/>
                <inhibitorKey/>
        </inhibitors>
</objectCharacteristics>
<creatingApplication>
        <creatingApplicationName>ESRI ArcCatalog</creatingApplicationName>
        <creatingApplicationVersion>9.1.0.722</creatingApplicationVersion>
        <dateCreatedByApplication>20050502</dateCreatedByApplication>
</creatingApplication>
<originalName>California.shp</originalName>
<storage>
        <contentLocation>
                <contentLocationType>URI</contentLocationType>
                <contentLocationValue>\\SUL-PM-JBANNING\NGDA\Data\ShapeFiles</contentLocationValue>
        </contentLocation>
        <storageMedium></storageMedium>
</storage>
<environment>
        <environmentCharacteristics>known to work</environmentCharacteristics>
        <environmentPurpose>edi/modify/render</environmentPurpose>
        <environmentNote/>
        <dependency>
                <dependencyName/>
                <dependencyIdentifier>
                        <dependencyIdentifierType/>
                        <dependencyIdentifierValue/>
                </dependencyIdentifier>
        </dependency>
        <software>
                <swName>ESRI ArcGIS</swName>
                <swVersion></swVersion>
                <swType>render</swType>
                <swOtherInformation/>
                <swDependency>Python 2.4</swDependency>
        </software>
```

```
		<hardware>
			<hwName>Intel Pentium II</hwName>
			<hwType>processor</hwType>
			<hwOtherInformation>Memory: 512 MB RAM</hwOtherInformation>
			<hwOtherInformation>Processor 1 GHz</hwOtherInformation>
		</hardware>
</environment>
<relationship>
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<!--dbf file-->
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>dbf_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--shx file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>shx_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--xml file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>xml_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--sbn file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>sbn_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--sbx file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>sbx_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--prj file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>prj_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
<!--Reresentation -->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>is child of</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>shapeFileAll_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
</relationship>
</ROOT>
```

## 2. shx file

```
<?xml version="1.0" encoding="UTF-8"?>
<ROOT>
<objectIdentifier>
        <objectIdentifierType>SDR_</objectIdentifierType>
        <objectIdentifierValue>shx_07108e3d-5fd1-11da-b211-19e7a5cf4814</objectIdentifierValue>
</objectIdentifier>
<preservationLevel>file</preservationLevel>
<objectCategory/>
<objectCharacteristics>
        <compositionLevel>0</compositionLevel>
        <fixity>
                <messageDigestAlgorithm>MD5</messageDigestAlgorithm>
                <messageDigest>bde9e5cbe7fdd652ea6d50734ae55f91</messageDigest>
                <messageDigestOriginator>Stanford Digital Repository</messageDigestOriginator>
        </fixity>
        <size>564</size>
        <format>
                <formatDesignation>
                        <formatName>Shapefile Index</formatName>
                        <formatVersion>1.0</formatVersion>
                </formatDesignation>
                <formatRegistry>
                 <!--NOTE: This is a placeholder as this format registry does not yet exist.-->
                        <formatRegistryName>NGDA Format Registry</formatRegistryName>
                        <formatRegistryKey>http://www.ngda.org/format/def/shapefile/shapefileIndex.tml</formatRegistryKey>
                        <formatRegistryRole>Specification</formatRegistryRole>
                </formatRegistry>
        </format>
        <significantProperties/>
        <inhibitors>
                <inhibitorType/>
                <inhibitorTarget/>
                <inhibitorKey/>
        </inhibitors>
</objectCharacteristics>
<creatingApplication>
        <creatingApplicationName>ESRI ArcCatalog</creatingApplicationName>
        <creatingApplicationVersion>9.1.0.722</creatingApplicationVersion>
        <dateCreatedByApplication>20050502</dateCreatedByApplication>
</creatingApplication>
<originalName>California.shx</originalName>
<storage>
        <contentLocation>
                <contentLocationType>URI</contentLocationType>
                <contentLocationValue>\\SUL-PM-JBANNING\NGDA\Data\ShapeFiles</contentLocationValue>
        </contentLocation>
        <storageMedium/>
</storage>
<environment>
        <environmentCharacteristics>known to work</environmentCharacteristics>
        <environmentPurpose>edi/modify/render</environmentPurpose>
        <environmentNote/>
        <dependency>
                <dependencyName/>
                <dependencyIdentifier>
                        <dependencyIdentifierType/>
                        <dependencyIdentifierValue/>
                </dependencyIdentifier>
        </dependency>
        <software>
                <swName>ESRI ArcGIS</swName>
                <swVersion/>
                <swType>render</swType>
                <swOtherInformation/>
                <swDependency>Python 2.4</swDependency>
        </software>
        <hardware>
                <hwName>Intel Pentium II</hwName>
```

```
                    <hwType>processor</hwType>
                    <hwOtherInformation>Memory: 512 MB RAM</hwOtherInformation>
                    <hwOtherInformation>Processor 1 GHz</hwOtherInformation>
            </hardware>
</environment>
<relationship>
            <relationshipType>structural</relationshipType>
            <relationshipSubType>has sibling</relationshipSubType>
            <!--dbf file-->
            <relatedObjectIdentification>
                    <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                    <relatedObjectIdentifierValue>dbf_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                    <relatedObjectSequence>0</relatedObjectSequence>
            </relatedObjectIdentification>
            <!--sbn file-->
            <relationshipType>structural</relationshipType>
            <relationshipSubType>has sibling</relationshipSubType>
            <relatedObjectIdentification>
                    <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                    <relatedObjectIdentifierValue>sbn_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                    <relatedObjectSequence>0</relatedObjectSequence>
            </relatedObjectIdentification>
            <!--xml file-->
            <relationshipType>structural</relationshipType>
            <relationshipSubType>has sibling</relationshipSubType>
            <relatedObjectIdentification>
                    <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                    <relatedObjectIdentifierValue>xml_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                    <relatedObjectSequence>0</relatedObjectSequence>
            </relatedObjectIdentification>
            <!--shp file-->
            <relationshipType>structural</relationshipType>
            <relationshipSubType>has sibling</relationshipSubType>
            <relatedObjectIdentification>
                    <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                    <relatedObjectIdentifierValue>shp_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                    <relatedObjectSequence>0</relatedObjectSequence>
            </relatedObjectIdentification>
            <!--sbx file-->
            <relationshipType>structural</relationshipType>
            <relationshipSubType>has sibling</relationshipSubType>
            <relatedObjectIdentification>
                    <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                    <relatedObjectIdentifierValue>sbx_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                    <relatedObjectSequence>0</relatedObjectSequence>
            </relatedObjectIdentification>
            <!--prj file-->
            <relationshipType>structural</relationshipType>
            <relationshipSubType>has sibling</relationshipSubType>
            <relatedObjectIdentification>
                    <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                    <relatedObjectIdentifierValue>prj_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                    <relatedObjectSequence>0</relatedObjectSequence>
            </relatedObjectIdentification>
            <!--Reresentation -->
            <relationshipType>structural</relationshipType>
            <relationshipSubType>is child of</relationshipSubType>
            <relatedObjectIdentification>
                    <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                    <relatedObjectIdentifierValue>shapeFileAll_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                    <relatedObjectSequence>0</relatedObjectSequence>
            </relatedObjectIdentification>
</relationship>
</ROOT>
```

## 3. dbf file

```
<?xml version="1.0" encoding="UTF-8"?>
<ROOT>
<objectIdentifier>
```

```
		<objectIdentifierType>SDR_</objectIdentifierType>
		<objectIdentifierValue>dbf_07108e3d-5fd1-11da-b211-19e7a5cf4814</objectIdentifierValue>
</objectIdentifier>
<preservationLevel>file</preservationLevel>
<objectCategory/>
<objectCharacteristics>
		<compositionLevel>0</compositionLevel>
		<fixity>
			<messageDigestAlgorithm>MD5</messageDigestAlgorithm>
			<messageDigest>d5ffe0573c1a1d2abe200c3cbf183efd</messageDigest>
			<messageDigestOriginator>Stanford Digital Repository</messageDigestOriginator>
		</fixity>
		<size>26232 </size>
		<format>
			<formatDesignation>
				<formatName>DBase</formatName>
				<formatVersion>1.0</formatVersion>
			</formatDesignation>
			<formatRegistry>
			 <!--NOTE: This is a placeholder as this format registry does not yet exist.-->
				<formatRegistryName>NGDA Format Registry</formatRegistryName>
				<formatRegistryKey>http://www.ngda.org/format/def/shapefile/DBase.html</formatRegistryKey>
				<formatRegistryRole>Specification</formatRegistryRole>
			</formatRegistry>
		</format>
		<significantProperties/>
		<inhibitors>
			<inhibitorType/>
			<inhibitorTarget/>
			<inhibitorKey/>
		</inhibitors>
</objectCharacteristics>
<creatingApplication>
		<creatingApplicationName>ESRI ArcCatalog</creatingApplicationName>
		<creatingApplicationVersion>9.1.0.722</creatingApplicationVersion>
		<dateCreatedByApplication>20050502</dateCreatedByApplication>
</creatingApplication>
<originalName>California.dbf</originalName>
<storage>
		<contentLocation>
			<contentLocationType>URI</contentLocationType>
			<contentLocationValue>\\SUL-PM-JBANNING\NGDA\Data\ShapeFiles</contentLocationValue>
		</contentLocation>
		<storageMedium/>
</storage>
<environment>
		<environmentCharacteristics>known to work</environmentCharacteristics>
		<environmentPurpose>edit/modify/render</environmentPurpose>
		<environmentNote/>
		<dependency>
			<dependencyName/>
			<dependencyIdentifier>
				<dependencyIdentifierType/>
				<dependencyIdentifierValue/>
			</dependencyIdentifier>
		</dependency>
		<software>
			<swName>ESRI ArcGIS</swName>
			<swVersion></swVersion>
			<swType>edit/modify/render</swType>
			<swOtherInformation/>
			<swDependency>Python 2.4</swDependency>
		</software>
		<hardware>
			<hwName>Intel Pentium II</hwName>
			<hwType>processor</hwType>
			<hwOtherInformation>Memory: 512 MB RAM</hwOtherInformation>
			<hwOtherInformation>Processor 1 GHz</hwOtherInformation>
		</hardware>
</environment>
```

```
<relationship>
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<!--sbn file-->
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>sbn_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--shx file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>shx_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--xml file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>xml_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--shp file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>shp_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--sbx file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>sbx_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--prj file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>prj_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--Reresentation -->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>is child of</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>shapeFileAll_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
</relationship>
</ROOT>
```

## 4. shp.xml file

```
<?xml version="1.0" encoding="UTF-8"?>
<ROOT>
<objectIdentifier>
		<objectIdentifierType>SDR_</objectIdentifierType>
		<objectIdentifierValue>xml_07108e3d-5fd1-11da-b211-19e7a5cf4814</objectIdentifierValue>
</objectIdentifier>
<preservationLevel>file</preservationLevel>
<objectCategory/>
```

```
<objectCharacteristics>
        <compositionLevel>0</compositionLevel>
        <fixity>
                <messageDigestAlgorithm>MD5</messageDigestAlgorithm>
                <messageDigest>37af5be9313b0c63de207bbc2622fa3a</messageDigest>
                <messageDigestOriginator>Stanford Digital Repository</messageDigestOriginator>
        </fixity>
        <size>126374</size>
        <format>
                <formatDesignation>
                        <formatName>XML</formatName>
                        <formatVersion>1.0</formatVersion>
                </formatDesignation>
                <formatRegistry>
                 <!--NOTE: This is a placeholder as this format registry does not yet exist.-->
                        <formatRegistryName>NGDA Format Registry</formatRegistryName>
                        <formatRegistryKey>http://www.ngda.org/format/def/shapefile/xml.html</formatRegistryKey>
                        <formatRegistryRole>Specification</formatRegistryRole>
                </formatRegistry>
        </format>
        <significantProperties/>
        <inhibitors>
                <inhibitorType/>
                <inhibitorTarget/>
                <inhibitorKey/>
        </inhibitors>
</objectCharacteristics>
<creatingApplication>
        <creatingApplicationName>ESRI ArcCatalog</creatingApplicationName>
        <creatingApplicationVersion>9.1.0.722</creatingApplicationVersion>
        <dateCreatedByApplication>20050502</dateCreatedByApplication>
</creatingApplication>
<originalName>California.shp.xml</originalName>
<storage>
        <contentLocation>
                <contentLocationType>URI</contentLocationType>
                <contentLocationValue>\\SUL-PM-JBANNING\NGDA\Data\ShapeFiles</contentLocationValue>
        </contentLocation>
        <storageMedium>hard-disc</storageMedium>
</storage>
<environment>
        <environmentCharacteristics>known to work</environmentCharacteristics>
        <environmentPurpose>edi/modify/render</environmentPurpose>
        <environmentNote/>
        <dependency>
                <dependencyName/>
                <dependencyIdentifier>
                        <dependencyIdentifierType/>
                        <dependencyIdentifierValue/>
                </dependencyIdentifier>
        </dependency>
        <software>
                <swName>ESRI ArcGIS</swName>
                <swVersion></swVersion>
                <swType>render</swType>
                <swOtherInformation/>
                <swDependency>Python 2.4</swDependency>
        </software>
        <hardware>
                <hwName>Intel Pentium II</hwName>
                <hwType>processor</hwType>
                <hwOtherInformation>Memory: 512 MB RAM</hwOtherInformation>
                <hwOtherInformation>Processor 1 GHz</hwOtherInformation>
        </hardware>
</environment>
<relationship>
        <relationshipType>structural</relationshipType>
        <relationshipSubType>has sibling</relationshipSubType>
        <!--dbf file-->
        <relatedObjectIdentification>
```

```
                <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                <relatedObjectIdentifierValue>dbf_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                <relatedObjectSequence>0</relatedObjectSequence>
        </relatedObjectIdentification>
        <!--shx file-->
        <relationshipType>structural</relationshipType>
        <relationshipSubType>has sibling</relationshipSubType>
        <relatedObjectIdentification>
                <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                <relatedObjectIdentifierValue>shx_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                <relatedObjectSequence>0</relatedObjectSequence>
        </relatedObjectIdentification>
        <!--sbn file-->
        <relationshipType>structural</relationshipType>
        <relationshipSubType>has sibling</relationshipSubType>
        <relatedObjectIdentification>
                <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                <relatedObjectIdentifierValue>sbn_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                <relatedObjectSequence>0</relatedObjectSequence>
        </relatedObjectIdentification>
        <!--shp file-->
        <relationshipType>structural</relationshipType>
        <relationshipSubType>has sibling</relationshipSubType>
        <relatedObjectIdentification>
                <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                <relatedObjectIdentifierValue>shp_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                <relatedObjectSequence>0</relatedObjectSequence>
        </relatedObjectIdentification>
        <!--sbx file-->
        <relationshipType>structural</relationshipType>
        <relationshipSubType>has sibling</relationshipSubType>
        <relatedObjectIdentification>
                <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                <relatedObjectIdentifierValue>sbx_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                <relatedObjectSequence>0</relatedObjectSequence>
        </relatedObjectIdentification>
        <!--prj file-->
        <relationshipType>structural</relationshipType>
        <relationshipSubType>has sibling</relationshipSubType>
        <relatedObjectIdentification>
                <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                <relatedObjectIdentifierValue>prj_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                <relatedObjectSequence>0</relatedObjectSequence>
        </relatedObjectIdentification>
        <!--Reresentation -->
        <relationshipType>structural</relationshipType>
        <relationshipSubType>is child of</relationshipSubType>
        <relatedObjectIdentification>
                <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                <relatedObjectIdentifierValue>shapeFileAll_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                <relatedObjectSequence>0</relatedObjectSequence>
        </relatedObjectIdentification>
</relationship>
</ROOT>
```

## 5. sbn file

```
<?xml version="1.0" encoding="UTF-8"?>
<ROOT>
<objectIdentifier>
        <objectIdentifierType>SDR_</objectIdentifierType>
        <objectIdentifierValue>sbn_07108e3d-5fd1-11da-b211-19e7a5cf4814</objectIdentifierValue>
</objectIdentifier>
<preservationLevel/>
<objectCategory>file</objectCategory>
<objectCharacteristics>
        <compositionLevel>0</compositionLevel>
        <fixity>
                <messageDigestAlgorithm>MD5</messageDigestAlgorithm>
```

```
                <messageDigest>63d3fc9d440fda99611863d7e81bddb3</messageDigest>
                <messageDigestOriginator>Stanford Digital Repository</messageDigestOriginator>
        </fixity>
        <size>732</size>
        <format>
                <formatDesignation>
                        <formatName>spatial Index</formatName>
                        <formatVersion>1.0</formatVersion>
                </formatDesignation>
                <formatRegistry>
                 <!--NOTE: This is a placeholder as this format registry does not yet exist.-->
                        <formatRegistryName>NGDA Format Registry</formatRegistryName>
                        <formatRegistryKey>http://www.ngda.org/format/def/shapefile/spatial_Index_SBN.html</formatRegistryKey>
                        <formatRegistryRole>Specification</formatRegistryRole>
                </formatRegistry>
        </format>
        <significantProperties/>
        <inhibitors>
                <inhibitorType/>
                <inhibitorTarget/>
                <inhibitorKey/>
        </inhibitors>
</objectCharacteristics>
<creatingApplication>
        <creatingApplicationName>ESRI ArcCatalog</creatingApplicationName>
        <creatingApplicationVersion>9.1.0.722</creatingApplicationVersion>
        <dateCreatedByApplication>20050502</dateCreatedByApplication>
</creatingApplication>
<originalName>California.sbn</originalName>
<storage>
        <contentLocation>
                <contentLocationType>URI</contentLocationType>
                <contentLocationValue>\\SUL-PM-JBANNING\NGDA\Data\ShapeFiles</contentLocationValue>
        </contentLocation>
        <storageMedium></storageMedium>
</storage>
<environment>
        <environmentCharacteristics>known to work</environmentCharacteristics>
        <environmentPurpose>edi/modify/render</environmentPurpose>
        <environmentNote/>
        <dependency>
                <dependencyName/>
                <dependencyIdentifier>
                        <dependencyIdentifierType/>
                        <dependencyIdentifierValue/>
                </dependencyIdentifier>
        </dependency>
        <software>
                <swName>ESRI ArcGIS</swName>
                <swVersion></swVersion>
                <swType>render</swType>
                <swOtherInformation/>
                <swDependency>Python 2.4</swDependency>
        </software>
        <hardware>
                <hwName>Intel Pentium II</hwName>
                <hwType>processor</hwType>
                <hwOtherInformation>Memory: 512 MB RAM</hwOtherInformation>
                <hwOtherInformation>Processor 1 GHz</hwOtherInformation>
        </hardware>
</environment>
<relationship>
        <relationshipType>structural</relationshipType>
        <relationshipSubType>has sibling</relationshipSubType>
        <!--dbf file-->
        <relatedObjectIdentification>
                <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                <relatedObjectIdentifierValue>dbf_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                <relatedObjectSequence>0</relatedObjectSequence>
        </relatedObjectIdentification>
```

```
		<!--shx file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>shx_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--xml file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>xml_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--shp file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>shp_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--sbx file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>sbx_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--prj file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>prj_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--Reresentation -->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>is child of</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>shapeFileAll_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
	</relationship>
	</ROOT>
```

## 6. sbx file

```
<?xml version="1.0" encoding="UTF-8"?>
<ROOT>
<objectIdentifier>
		<objectIdentifierType>SDR_</objectIdentifierType>
		<objectIdentifierValue>sbx_07108e3d-5fd1-11da-b211-19e7a5cf4814</objectIdentifierValue>
</objectIdentifier>
<preservationLevel/>
<objectCategory>file</objectCategory>
<objectCharacteristics>
		<compositionLevel>0</compositionLevel>
		<fixity>
			<messageDigestAlgorithm>MD5</messageDigestAlgorithm>
			<messageDigest>5de669348a10f2bfa73b623cf0b9167f</messageDigest>
			<messageDigestOriginator>Stanford Digital Repository</messageDigestOriginator>
		</fixity>
		<size>164</size>
		<format>
```

```
                        <formatDesignation>
                                <formatName>spatial Index</formatName>
                                <formatVersion>1.0</formatVersion>
                        </formatDesignation>
                        <formatRegistry>
                         <!--NOTE: This is a placeholder as this format registry does not yet exist.-->
                                <formatRegistryName>NGDA Format Registry</formatRegistryName>
                                <formatRegistryKey>http://www.ngda.org/format/def/shapefile/shape_index_SBX.html</formatRegistryKey>
                                <formatRegistryRole>Specification</formatRegistryRole>
                        </formatRegistry>
                </format>
                <significantProperties/>
                <inhibitors>
                        <inhibitorType/>
                        <inhibitorTarget/>
                        <inhibitorKey/>
                </inhibitors>
        </objectCharacteristics>
        <creatingApplication>
                <creatingApplicationName>ESRI ArcCatalog</creatingApplicationName>
                <creatingApplicationVersion>9.1.0.722</creatingApplicationVersion>
                <dateCreatedByApplication>20050502</dateCreatedByApplication>
        </creatingApplication>
        <originalName>California.sbx</originalName>
        <storage>
                <contentLocation>
                        <contentLocationType>URI</contentLocationType>
                        <contentLocationValue>\\SUL-PM-JBANNING\NGDA\Data\ShapeFiles</contentLocationValue>
                </contentLocation>
                <storageMedium></storageMedium>
        </storage>
        <environment>
                <environmentCharacteristics>known to work</environmentCharacteristics>
                <environmentPurpose>edi/modify/render</environmentPurpose>
                <environmentNote/>
                <dependency>
                        <dependencyName/>
                        <dependencyIdentifier>
                                <dependencyIdentifierType/>
                                <dependencyIdentifierValue/>
                        </dependencyIdentifier>
                </dependency>
                <software>
                        <swName>ESRI ArcGIS</swName>
                        <swVersion></swVersion>
                        <swType>render</swType>
                        <swOtherInformation/>
                        <swDependency>Python 2.4</swDependency>
                </software>
                <hardware>
                        <hwName>Intel Pentium II</hwName>
                        <hwType>processor</hwType>
                        <hwOtherInformation>Memory: 512 MB RAM</hwOtherInformation>
                        <hwOtherInformation>Processor 1 GHz</hwOtherInformation>
                </hardware>
        </environment>
        <relationship>
                <relationshipType>structural</relationshipType>
                <relationshipSubType>has sibling</relationshipSubType>
                <!--dbf file-->
                <relatedObjectIdentification>
                        <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                        <relatedObjectIdentifierValue>dbf_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                        <relatedObjectSequence>0</relatedObjectSequence>
                </relatedObjectIdentification>
                <!--shx file-->
                <relationshipType>structural</relationshipType>
                <relationshipSubType>has sibling</relationshipSubType>
                <relatedObjectIdentification>
                        <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
```

<related_object_identifier_value>shx_07108e3d-5fd1-11da-b211-19e7a5cf4814</related_object_identifier_value>
<related_object_sequence>0</related_object_sequence>
</related_object_identification>
<!--xml file-->
<relationship_type>structural</relationship_type>
<relationship_sub_type>has sibling</relationship_sub_type>
<related_object_identification>
<related_object_identifier_type>SDR_</related_object_identifier_type>
<related_object_identifier_value>xml_07108e3d-5fd1-11da-b211-19e7a5cf4814</related_object_identifier_value>
<related_object_sequence>0</related_object_sequence>
</related_object_identification>
<!--shp file-->
<relationship_type>structural</relationship_type>
<relationship_sub_type>has sibling</relationship_sub_type>
<related_object_identification>
<related_object_identifier_type>SDR_</related_object_identifier_type>
<related_object_identifier_value>shp_07108e3d-5fd1-11da-b211-19e7a5cf4814</related_object_identifier_value>
<related_object_sequence>0</related_object_sequence>
</related_object_identification>
<!--sbn file-->
<relationship_type>structural</relationship_type>
<relationship_sub_type>has sibling</relationship_sub_type>
<related_object_identification>
<related_object_identifier_type>SDR_</related_object_identifier_type>
<related_object_identifier_value>sbn_07108e3d-5fd1-11da-b211-19e7a5cf4814</related_object_identifier_value>
<related_object_sequence>0</related_object_sequence>
</related_object_identification>
<!--prj file-->
<relationship_type>structural</relationship_type>
<relationship_sub_type>has sibling</relationship_sub_type>
<related_object_identification>
<related_object_identifier_type>SDR_</related_object_identifier_type>
<related_object_identifier_value>prj_07108e3d-5fd1-11da-b211-19e7a5cf4814</related_object_identifier_value>
<related_object_sequence>0</related_object_sequence>
</related_object_identification>
<!--Reresentation -->
<relationship_type>structural</relationship_type>
<relationship_sub_type>is child of</relationship_sub_type>
<related_object_identification>
<related_object_identifier_type>SDR_</related_object_identifier_type>
<related_object_identifier_value>shapeFileAll_07108e3d-5fd1-11da-b211-19e7a5cf4814</related_object_identifier_value>
<related_object_sequence>0</related_object_sequence>
</related_object_identification>
</relationship>
</ROOT>

## 7. prj file

```
<?xml version="1.0" encoding="UTF-8"?>
<ROOT>
<objectIdentifier>
<objectIdentifierType>SDR_</objectIdentifierType>
<objectIdentifierValue>prj_07108e3d-5fd1-11da-b211-19e7a5cf4814</objectIdentifierValue>
</objectIdentifier>
<preservationLevel/>
<objectCategory>file</objectCategory>
<objectCharacteristics>
<compositionLevel>0</compositionLevel>
<fixity>
<messageDigestAlgorithm>MD5</messageDigestAlgorithm>
<messageDigest>8e24fe15b2c8c640c459006722fa1e7f</messageDigest>
<messageDigestOriginator>Stanford Digital Repository</messageDigestOriginator>
</fixity>
<size>167</size>
<format>
<formatDesignation>
<formatName>projection</formatName>
<formatVersion>1.0</formatVersion>
</formatDesignation>
<formatRegistry>
```

```
                        <!--NOTE: This is a placeholder as this format registry does not yet exist.-->
                                <formatRegistryName>NGDA Format Registry</formatRegistryName>
                                <formatRegistryKey>http://www.ngda.org/format/def/shapefile/projectionFile.html</formatRegistryKey>
                                <formatRegistryRole>Specification</formatRegistryRole>
                        </formatRegistry>
                </format>
                <significantProperties/>
                <inhibitors>
                        <inhibitorType/>
                        <inhibitorTarget/>
                        <inhibitorKey/>
                </inhibitors>
        </objectCharacteristics>
        <creatingApplication>
                <creatingApplicationName>ESRI ArcCatalog</creatingApplicationName>
                <creatingApplicationVersion>9.1.0.722</creatingApplicationVersion>
                <dateCreatedByApplication>20050502</dateCreatedByApplication>
        </creatingApplication>
        <originalName>California.prj</originalName>
        <storage>
                <contentLocation>
                        <contentLocationType>URI</contentLocationType>
                        <contentLocationValue>\\SUL-PM-JBANNING\NGDA\Data\ShapeFiles</contentLocationValue>
                </contentLocation>
                <storageMedium></storageMedium>
        </storage>
        <environment>
                <environmentCharacteristics>known to work</environmentCharacteristics>
                <environmentPurpose>edi/modify/render</environmentPurpose>
                <environmentNote/>
                <dependency>
                        <dependencyName/>
                        <dependencyIdentifier>
                                <dependencyIdentifierType/>
                                <dependencyIdentifierValue/>
                        </dependencyIdentifier>
                </dependency>
                <software>
                        <swName>ESRI ArcGIS</swName>
                        <swVersion></swVersion>
                        <swType>render</swType>
                        <swOtherInformation/>
                        <swDependency>Python 2.4</swDependency>
                </software>
                <hardware>
                        <hwName>Intel Pentium II</hwName>
                        <hwType>processor</hwType>
                        <hwOtherInformation>Memory: 512 MB RAM</hwOtherInformation>
                        <hwOtherInformation>Processor 1 GHz</hwOtherInformation>
                </hardware>
        </environment>
        <relationship>
                <relationshipType>structural</relationshipType>
                <relationshipSubType>has sibling</relationshipSubType>
                <!--dbf file-->
                <relatedObjectIdentification>
                        <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                        <relatedObjectIdentifierValue>dbf_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                        <relatedObjectSequence>0</relatedObjectSequence>
                </relatedObjectIdentification>
                <!--shx file-->
                <relationshipType>structural</relationshipType>
                <relationshipSubType>has sibling</relationshipSubType>
                <relatedObjectIdentification>
                        <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                        <relatedObjectIdentifierValue>shx_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                        <relatedObjectSequence>0</relatedObjectSequence>
                </relatedObjectIdentification>
                <!--xml file-->
                <relationshipType>structural</relationshipType>
```

```
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>xml_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--shp file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>shp_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--sbx file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>sbx_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--sbn file-->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>has sibling</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>sbn_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
		<!--Reresentation -->
		<relationshipType>structural</relationshipType>
		<relationshipSubType>is child of</relationshipSubType>
		<relatedObjectIdentification>
			<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
			<relatedObjectIdentifierValue>shapeFileAll_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
			<relatedObjectSequence>0</relatedObjectSequence>
		</relatedObjectIdentification>
</relationship>
</ROOT>
```

## 8. Conceptual Shapefile representation

```
<?xml version="1.0" encoding="UTF-8"?>
<ROOT>
<objectIdentifier>
		<objectIdentifierType>SDR_</objectIdentifierType>
		<objectIdentifierValue>shapeFileAll_07108e3d-5fd1-11da-b211-19e7a5cf4814</objectIdentifierValue>
</objectIdentifier>
<preservationLevel/>
<objectCategory>representation</objectCategory>
<objectCharacteristics>
		<compositionLevel/>
		<fixity>
			<messageDigestAlgorithm/>
			<messageDigest/>
			<messageDigestOriginator/>
		</fixity>
		<size/>
		<format>
			<formatDesignation>
				<formatName>ESRI Shapefile</formatName>
				<formatVersion>1.0</formatVersion>
			</formatDesignation>
			<formatRegistry>
			 <!--NOTE: This is a placeholder as this format registry does not yet exist.-->
				<formatRegistryName>NGDA Format Registry</formatRegistryName>
				<formatRegistryKey>http://www.ngda.org/format/def/shapefile/shapefile.html</formatRegistryKey>
```

```
                        <formatRegistryRole>Specification</formatRegistryRole>
                    </formatRegistry>
            </format>
            <significantProperties/>
            <inhibitors>
                    <inhibitorType/>
                    <inhibitorTarget/>
                    <inhibitorKey/>
            </inhibitors>
</objectCharacteristics>
<creatingApplication>
            <creatingApplicationName>ESRI ArcGIS</creatingApplicationName>
            <creatingApplicationVersion>9.1.0.722</creatingApplicationVersion>
            <dateCreatedByApplication>20050502</dateCreatedByApplication>
</creatingApplication>
<originalName>California.shp</originalName>
<storage>
            <contentLocation>
                    <contentLocationType>URI</contentLocationType>
                    <contentLocationValue>t </contentLocationValue>
            </contentLocation>
            <storageMedium/>
</storage>
<environment>
            <environmentCharacteristics/>
            <environmentPurpose/>
            <environmentNote/>
            <dependency>
                    <dependencyName/>
                    <dependencyIdentifier>
                            <dependencyIdentifierType/>
                            <dependencyIdentifierValue/>
                    </dependencyIdentifier>
            </dependency>
            <software>
                    <swName>ESRI ArcGIS </swName>
                    <swVersion>9.1.0.722</swVersion>
                    <swType>render</swType>
                    <swOtherInformation/>
                    <swDependency/>
            </software>
            <software>
                    <swName>Windows NT</swName>
                    <swVersion> 5.0 </swVersion>
                    <swType>operatingSystem </swType>
                    <swOtherInformation/>
                    <swDependency/>
            </software>
            <hardware>
                    <hwName>Intel Pentium II</hwName>
                    <hwType>processor</hwType>
                    <hwOtherInformation>Memory 512 MB RAM</hwOtherInformation>
                    <hwOtherInformation>Processor 1 GHz</hwOtherInformation>
            </hardware>
</environment>
<relationship>
            <relationshipType>structural</relationshipType>
            <relationshipSubType>has sibling</relationshipSubType>
            <!--dbf file-->
            <relatedObjectIdentification>
                    <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                    <relatedObjectIdentifierValue>dbf_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
                    <relatedObjectSequence>0</relatedObjectSequence>
            </relatedObjectIdentification>
            <!--shx file-->
            <relationshipType>structural</relationshipType>
            <relationshipSubType>has sibling</relationshipSubType>
            <relatedObjectIdentification>
                    <relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
                    <relatedObjectIdentifierValue>shx_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
```

```
		<relatedObjectSequence>0</relatedObjectSequence>
</relatedObjectIdentification>
<!--xml file-->
<relationshipType>structural</relationshipType>
<relationshipSubType>has sibling</relationshipSubType>
<relatedObjectIdentification>
		<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
		<relatedObjectIdentifierValue>xml_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
		<relatedObjectSequence>0</relatedObjectSequence>
</relatedObjectIdentification>
<!--shp file-->
<relationshipType>structural</relationshipType>
<relationshipSubType>has sibling</relationshipSubType>
<relatedObjectIdentification>
		<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
		<relatedObjectIdentifierValue>shp_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
		<relatedObjectSequence>0</relatedObjectSequence>
</relatedObjectIdentification>
<!--sbx file-->
<relationshipType>structural</relationshipType>
<relationshipSubType>has sibling</relationshipSubType>
<relatedObjectIdentification>
		<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
		<relatedObjectIdentifierValue>sbx_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
		<relatedObjectSequence>0</relatedObjectSequence>
</relatedObjectIdentification>
<!--prj file-->
<relationshipType>structural</relationshipType>
<relationshipSubType>has sibling</relationshipSubType>
<relatedObjectIdentification>
		<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
		<relatedObjectIdentifierValue>prj_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
		<relatedObjectSequence>0</relatedObjectSequence>
</relatedObjectIdentification>
<!--sbn file-->
<relationshipType>structural</relationshipType>
<relationshipSubType>has sibling</relationshipSubType>
<relatedObjectIdentification>
		<relatedObjectIdentifierType>SDR_</relatedObjectIdentifierType>
		<relatedObjectIdentifierValue>sbn_07108e3d-5fd1-11da-b211-19e7a5cf4814</relatedObjectIdentifierValue>
		<relatedObjectSequence>0</relatedObjectSequence>
</relatedObjectIdentification>
</relationship>
</ROOT>
```

## **Appendix B –**Lineage from Minnesota Land Use and Cover: 1990s Census of the Land (http://lucy.lmic.state.mn.us/metadata/luse8.html)

**Land Use Data Sources:**

Agricultural and Transition Areas
Forested Areas
Interpreted TM satellite imagery for the Twin Cities metro area
Generalized Land Use for the Twin Cities Metropolitan Area (only the farmstead category)
Olmsted County
Beltrami and Clearwater Counties
Camp Ripley and Beltrami Island State Forest

County Boundaries Data Source:
MnDNR's CTYBDNE2 coverage (see documentation at
http://deli.dnr.state.mn.us/metadata/full/ctybdne2.html )

DNR's Regional Boundaries Data Source:
DNR Regions coverage (see documentation at http://deli.dnr.state.mn.us/metadata/full/dnrrgne2.html )

**MnDNR's Processing Steps**:

All land use/cover data was put together by county in raster format using Arc/INFO GRIDs. The data that existed as vector data sets (Agriculture and Transition Areas, farmstead category from the Metropolitan Council data set, and Olmsted County) was rasterized to 30 meter by 30 meter cells prior to mosaicking using the THEME menu, Convert to Grid option in ArcView's Spatial Analyst. All county tiles were based on DNR's CTYBDNE2 coverage.

Special Processing for the metro area: Two data sets were used in the metro area. All land use classifications in the interpreted TM satellite imagery data set were used since they more closely matched classifications used in other areas in Minnesota. The one class that was not well-represented in the TM data set was scattered houses so the farmstead class from the Metropolitan Council land use data was incorporated into the TM data. This was done using simple overlay techniques in Spatial Analyst.

Individual county data sets were merged into tiles based on DNR's Administrative Regions. The DNR Administrative regions coverage was derived from the CTYBDNE2 coverage since most regional boundaries are based on county borders.

Each regional landuse/cover grid was then subjected to the following clean-up process. When raster data is mosaicked, there are gaps that occur between the tiles where they did not match up perfectly. Typically these gaps are very small, on the order of one or two cells in width. To fill in these gaps, the NIBBLE process in Spatial Analyst was used to replace cells that were offsite by using nearest neighbor rules. Each data set was masked so that only those cells within each region were processed. This is similar to a clip command in a vector GIS system.

Each of the regional data set grids were then mosaicked together using the MERGE request and then cleaned-up using the NIBBLE request as described above.

The resulting landuse/cover grid had one attribute called VALUE. This item contained the attribute codes for each of the different landuse/cover classes from each of the differing coding schemes. Since there were 6 sources for the data and since there were 5 different coding schemes, a new coding scheme had to be developed to maintain data integrity. To accomplish this, the data from different sources was offset in the following fashion:

100 Beltrami / Clearwater Counties
200 Camp Ripley / Beltrami Island State Forest

300 Forested
400 Olmsted County
500 Ag and Transition Areas
600 Twin Cities metro (TM and farmsteads)

Using this coding scheme, every unique data value was preserved. In all but Olmsted County, the data sets were simply offset by the appropriate value. For Olmsted County, where the landuse and cover class values exceeded 100, they were simply numbered sequentially from 1 to 37 and then offset by 400.

A lookup table (lulookup.dbf) was then created with the following fields:
New_code - The new code as it exists in the statewide grid
Orig_code - The Original code as it existed in the source data
Map_code - The codes as they were assigned on the statewide 1990s Land Use and Cover map
Orig_desc - The Original class description
Map_desc - The Class descriptions as shown on the statewide 1990s Land Use and Cover map

This table could be related/joined to the grid table using the VALUE item in the GRID and the NEW_CODE item in the lookup table.

Files for Public Distribution: A file that contained only the NEW_CODE item was created for public distribution. It is available in ArcGRID and EPPL7 raster formats. The lookup table, lulookup.dbf, is provided to show how the detailed legend categories in the original data sets were matched to one of the eight land use categories in this data set.

Several reported errors were corrected (4/2000):
1. City of Roseau: the western portion of the city was recoded from cultivated (2) to urban (1).
2. Chisago County: two small areas along the northern county boundary were recoded from forested (5) to unknown (9).
3. City of Wabasha: the northern portion of the city was recoded from water (6) to urban (1).
4. City of Hammond: the eastern portion of the city was recoded from cultivated (2) to urban (1).
5. Olmsted County: an area just northeast of the city of Rochester was recoded from unknown (9) to cultivated (2).

**Appendix C: Retention and Storage of Technical Characteristics of a Shapefile**

**Introduction:**
It is not enough to capture specific characteristics associated with a preservation data object or file (i.e. environment, computing platform, file size, file relationships, provider, etc) to fully understand the object. Additional documentation, such as relevant specifications and related source materials, are needed to fully understand the appropriate use and context of the data type or format. Since this kind of information is not specific to an instance of the data type or format, many organizations such as the PREMIS Working Group have contended that such information common to the data types or formats should be kept in one place and managed by an authoritative source. For example, the purpose of a format registry should be to answer questions such as how should a Digital Raster Graphic or shapefile be used? Or what is the wave length range for a band 3 LandSat7 scene? A format registry should also address questions such as what additional information is necessary to comprehend a given data type/format before attempted use.

As stated previously, the PREMIS data model relies on the use of a format registry to contain information at a higher level rather than store it for each individual digital object in an archive. The role of the format registry in PREMIS is a location to discover additional characteristics intrinsic to any given entry. For instance, upon obtaining GIS data complete with FGDC metadata from an archive, it would be helpful to a user to be able to have access to the content standard to understand the element definitions. In the future, without a reference to the content standard, how will the following tags be deciphered?

```
<attr>
        <attrlabl Sync="TRUE">AVG_SALE87</attrlabl>
        <attalias Sync="TRUE">AVG_SALE87</attalias>
        <attrtype Sync="TRUE">Number</attrtype>
        <attwidth Sync="TRUE">7</attwidth>
</attr>
```

**Case Study:**
The following is an exercise in looking at a common geospatial data types/format and examining the technical characteristics and other information necessary to archive it. Suggestions are made about where this kind of information should be stored, i.e., in a format registry or in a submission package to a given archive or repository?

**ESRI Shapefile Case Study:**
**How do you preserve a shapefile?**

**What is a shapefile?**

Originally developed by ESRI to work with their ArcView application, shapefiles have become one of the most widely used and recognized geospatial vector data types today. According to the ESRI specification, "a shapefile consists of a main file, an index file, and a dBASE table" all with the same name prefix (i.e. states.shp, states.shx, states.dbf). The .shp file, also known as the main file,

contains multiple records which describes a shape with a list of vertices. This file stores spatial geometry for features. The related index file (.shx) contains the "offset of the corresponding main file record from the beginning of the main file". The dBase file (.dbf) contains feature attributes, where each feature is related to one entry in the dBase table. Image X shows a rendered shapefile of county boundaries in California.

Image X. Rendered shapefile of California counties.

Additional files may supplement the three core files that comprise the ESRI shapefile data type. More common supplementary files include projection files (*.prj) which store spatial coordinate information, spatial index for the geometric data (*.sbn and *.sbx), and metadata files (*.shp.xml) which contain descriptive and technical information about the shapefile as a whole. Shapefiles are also flexible and support joining additional tables to the original .dbf file. This extends the attributes that can be related to spatial features. An example of this is joining census data to the .dbf file using the zip code field as the primary and foreign keys. For a more complete understanding, the ESRI shapefile Technical Description is available from the following website: http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf.

**ESRI shapefile Preservation**

Upon obtaining a representative shapefile from the California Spatial Information Library (CaSEIL) an investigation revealed that seven files comprise this particular shapefile.

- The main file (.shp)
- An index file (.shx)
- Database file (.dbf)
- Projection file (.prj)

- Spatial index files (.sbn. and .sbx)
- Metadata file (.shp.xml)

As stated above and in the ESRI shapefile Technical Description, there are only three *core* files (.shp, .dbf, ,shx) needed to make up a shapefile. It is therefore understood that these must be retained for preservation. Additionally the projection file (.prj) is a text file which defines the map projection of the coordinates in the shapefile and should be preserved when available. The other files (.sbn, .sbx, .shp.xml) can be considered contextual and may be retained dependent upon an institution's preservation policy. Arguably in this case, ignoring or deleting the optional files would result in loss in the understanding of the data as explained below.

The shp.xml file contains an FGDC metadata record for the data set. This content standard has fields for providing comments on the fitness of the data, the appropriate uses of the data, as well as use constraints. Additional fields provide the opportunity for detailed attribute definitions that may be codes in the shapefile attribute table.

The last two files present are optional spatial index files (.sbn, .sbx). These files are "used to improve access performance in some applications" but are not necessary for rendering or editing. The index files and the presence of spatial indices may be interpreted as contextual information for the dataset. For instance, a spatial index may be a commentary on the data complexity. Since there are a significant number of geographic features (points, arcs, polygons), one could speculate that a spatial index is provided to aid in performance. To more authoritatively ascertain how and when these spatial index files get created, additional investigation is needed.

A preservation policy would ultimately determine which files contributing to a shapefile are kept for preservation. The minimum requirement as detailed in the technical specification are the .dbf, .shp, and .shx, but an argument can be made that additional information, both contextual and appropriate usage, can be gleaned from the optional files accompanying the core files. This makes the optional files valuable in terms of preservation. Also, ignoring or removing them might prove to be more trouble than including them. Appendix A illustrates how the PREMIS scheme can be used to document all the files contained in the shapefile.

**What should be in the format registry for ESRI Shapefile?**

ESRI has written a technical specification on the shapefile data type that must be considered the authoritative source. While the paper contains detailed information on the main files that make up a shapefile (.shp, .shx, .dbf), there is no discussion of the other file types that may be included. As we have seen in the shapefile data type obtained for CaSIL, additional files may exist in that are not mentioned in technical specification. Ideally, documentation and information on those files would also be contained in the ESRI Shapefile entry of a format registry.

Documentation to be included in the ESRI shapefile format registry:

1. ESRI shapefile Technical Description - http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf
2. dBase specification – .dbf files are one of the core shapefile components. dBase files are also used when joining attribute tables. (http://www.clicketyclick.dk/databases/xbase/format/)
3. Ideally, additional documentation, specifications or statements on the various files that may be used as part of shapefiles although no known publication exists detailing what files may be part of a shapefile data type. An investigation concluded that .sbn, .sbx, .prj, . xml, .fbn, .fbx, .ain and .aih files may all be included in a shapefile data type. Documentation as well as a through understanding of the roles/purpose of these files is also not available on the above mentioned file types.
4. Specifications for the different geospatial metadata standards (FGDC, ANZLIC, CEN, etc.) referenced by the optional metadata file provided.

**What about incomplete data formats specifications?**
Inconsistencies between data type specifications and the actual files found in a digital object exist as was obvious when investigating ESRI shapefiles obtained from CaSIL. For the investigated shapefiles, numerous files that comprised the shapefile data type (.sbx, .sbn, .prj, .shp.xml) were not included in ESRI's technical documentation (http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf). After discussing this with ESRI, a comprehensive list of the possible files which may be included as part of the shapefile data type as well as their role was provided. It should be noted that this list is not included in any technical specification or official white paper, but was only available after posing the question to the ESRI technical support staff.

| File Extension | File Role | Contained in specification |
|---|---|---|
| ain | attribute index file | |
| aih | attribute index file | |
| dbf | Shapefile attribute table file | Y |
| fbn | spatial index file for read | |
| fbx | spatial index file for read | |
| idx | geocoding index for read | |
| ixs | geocoding index for read | |
| mxs | geocoding index for read | |
| prj | projections definition file | |
| sbn | spatial index for read | Y |
| sbx | spatial index for read | Y |

While other data types (DRG, DOQ, Landsat 7) obtained from CaSIL for analysis all contained the *minimum* file requirements as detailed in the specification, it was often necessary to capture additional files and contextual information to completely understand the data.

To get the most value from metadata, the specification or standard to which the document adheres should be referenced and available in the format registry. Typically, metadata documents are available in some form of mark-up language and represent an instance of the specification. The metadata is difficult to decipher without knowing what the elements represent, or providing a means to discover them; thus, the inclusion of the collection of relevant geospatial metadata standards would be wise.

**CONCLUSIONS:**

With the advancements in technology, documenting geospatial datasets is becoming easier and less burdensome for GIS professionals. Several of the major GIS software vendors (ESRI, Intergraph) have brought metadata to the forefront by providing metadata editors as part of the core application. "Sychronizers", i.e., software code that can capture specific characteristics of the data set and maintain them in a metadata document, are also commonly included in GIS software packages. Customization of both synchronizers and editors allows flexibility in determining which details of a data set to capture. This emphasis upon metadata by software companies coincides with the Federal government's initiative to promote geospatial data, as highlighted in the GeoSpatial One Stop activities (http://www.geo-one-stop.gov/). More importantly all of these activities lead to a wider metadata user base and a general education on metadata throughout the GIS community. Those involved with geospatial data are more aware than ever of the importance of well documented data. A common terminology is emerging that allows professionals to speak to each other about data set characteristics (quality, access and use restrictions, spatial reference information, entity and attribute information, etc.).

In terms of preservation, the importance of including metadata with geospatial data is becoming more clear. As discussed, many of the elements contained in the FGDC content standard and subsequent community profiles relate to preservation concepts (environment, computing platform, semantics, domain specific terminology, provenance, provider, quality, and appropriate use). The cost of including such a metadata record in a preservation repository, *when already available*, is close to nothing. Technological advancements in the metadata tools have helped to drive the costs down of creating such metadata, yet they are far from insignificant for data creators or publishers. More research needs to be done to show that the benefits of having the information on a data set's characteristics outweigh the costs.