Utilizing Geospatial Metadata to Support Data Preservation Practices

Background:

Metadata associated with geospatial datasets can provide a rich insight into the technical details about the dataset it is describing while also providing information about the 'who', 'what', 'when', 'where', 'why' and 'how' to explain the dataset's purpose and utility. When thoughtfully populated, geospatial metadata can be a critical resource for understanding and managing geospatial data for current and future GIS practitioners and those trying to preserve the data.

The two primary geospatial metadata standards explored by GeoMAPP in its metadata explorations are the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) - FGDC-STD-001-1998,[1] and the International Organization for Standardization (ISO) - 19115:2003 Standard for Geographic Information Metadata.[2] The team investigated in detail the FGDC CSDGM's domain of metadata fields and mandatory elements. Due to the standard's prominence and ubiquity in the geospatial community the team chose to follow this metadata schema and its recommended required fields as the project standard for geospatial metadata. The team also utilized the ISO standard's 19 Topical Categories, which range from Biota and Boundaries to Structures and Utilities, to help group, organize and manage their data. Each state has their geoarchival holdings organized within ISO Topic Categories.

Within the GeoMAPP partners, metadata plays a large role in each state's geospatial data clearinghouse and is increasing in importance for archiving activities. Each state requires that metadata meeting the FGDC's CSDGM must be included with any dataset that is to be hosted on the state's data clearinghouse. In North Carolina and Utah, geospatial metadata is also reviewed by archives staff when data is transferred and ingested into state geoarchival repositories. As part of the project's activities, the team has also compared and cross-walked the FGDC CSDGM with the more archives-centric Dublin Core Metadata Standard.[3] The North Carolina team has used this research to manually extract key fields from geospatial metadata to populate their MARS online archives catalog[4] and ContentDM digital collections access solution.[5] Kentucky also experimented with mapping geospatial metadata to corresponding Dublin Core fields in its DSpace data repository. The Utah team is working on developing a metadata parser to automate the integration of geospatial metadata elements into their AXAEM archives management tool.

Despite metadata's importance for data management and understanding, metadata creation and maintenance is often considered to be a "nice to have" for resource constrained geospatial data creators and is often not part of the critical path for their data creation workflows. As a result, metadata is either incomplete or is missing entirely. Each of the GeoMAPP partners require that every dataset submitted for inclusion in the state clearinghouse has an accompanying FGDC CSDGM compliant metadata record. However, the reality is that the clearinghouse GIS staff often has to work with the data producers directly to help create these records, or will create the metadata

[1] FGDC Content Standard for Geospatial Metadata http://www.fgdc.gov/metadata/geospatial-metadata-standards#csdgm
Graphical View of FGDC Content Standard for Geospatial Metadata http://www.fgdc.gov/csdgmgraphical/index.html
[2] ISO 19115:2003 http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020
[3] Dublin Core Metadata: http://dublincore.org/
[4] NC MARS Catalog: http://mars.archives.ncdcr.gov/
[5] NC ContentDM Access solutions: http://digital.ncdcr.gov/cdm4/additional_collections.php

record from scratch for the data producer. The Kentucky team stores important datasets submitted to them with lacking or incomplete metadata in a separate "conditional" database; a data purgatory where data resides waiting for a competed metadata record before it can be published. Limitations of metadata validation tools further compound the challenge of reviewing and managing metadata. Due to the length and richness of geospatial metadata and the time consuming nature of manually reviewing individual records, metadata is also often validated with automated parsing tools that check to see that required fields are populated. However, the parsers have no way of checking the quality and completeness of the metadata field values.[6]

Given these limitations and challenges and the increasing value of geospatial metadata for preservation and future access to the GIS datasets, the following tables identify and describe key geospatial metadata fields that can be beneficial for long term preservation, and enabling access to superseded geospatial data. While fully compliant, richly completed full geospatial metadata records should always be the preferred standard for GIS data creators, the following lists highlight metadata fields that deserve special focus to be thoughtfully populated by GIS data creators and more thoroughly reviewed by GIS clearinghouses and archives to benefit GIS data preservation, access, and use.

CSDGM Checklist

The FGDC-STD-001-1998 organizes the geospatial metadata elements into seven sections. The metadata is further organized into a hierarchy of data elements and compound data elements that define the information content for the metadata to document a set of digital geospatial data.[7]

1. Identification Information - basic information about the data set

2. Data Quality Information - provides a general assessment of the quality of the data set.

3. Spatial Data Organization Information - the mechanism used to represent spatial information in the data set

4. Spatial Reference Information - the description of the reference frame for, and the means to encode, coordinates in the data set

5. Entity and Attribute Information - details about the information content of the data set, the entities, their attributes, and domains from which attribute values may be assigned

6. Distribution Information - information about the distributor and options for obtaining the data set

7. Metadata Reference Information - information on the part responsible for creating the metadata and the currentness of the metadata

The following table offers a checklist of important CSDGM fields that will facilitate long-term preservation of the geospatial datasets. The checklist is organized based on the FDGC-STD-001-1998 standard. **Bold items** are metadata fields that should be provided by the geospatial data producer. Note: several of these items may be automatically populated by the GIS software. These have been denoted with a green asterisk: *.

[6] USGS Geospatial (FGDC) Metadata Validation Service http://geo-nsdi.er.usgs.gov/validation/

[7] FGDC-STD-001-1998 (pg. vii): http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf

Checklist: CSDGM GIS Metadata Fields for Preservation

Important CSDGM Fields for Preservation:

	Friendly Field Name	XML Tag	Description / Example Value
1. Identification Information Section		<idinfo>	
1.1 Citation 1.1.8 Citation Information		<citation> <citeinfo>	
	Originators	**<origin>**	The **party responsible** for the dataset. This is often the dataset creator except in cases where the dataset's creation was contracted out to a third party, but is 'owned' and maintained by another party.
	Publication Date	**<pubdate>**	The date the dataset was **completed** and was made ready for **use**.
	Title	**<title>**	Title of dataset, ideally including 'where', 'what', 'when' *e.g. North Carolina Shellfish Growing Areas 2010*
*	**Geospatial Data Presentation Form**	**<geoform>**	Describes type and format of the dataset *e.g. vector digital data*
1.1.8.8 Publication Information		<pubinfo>	
	Publication Place	**<pubplace>**	The place/location the dataset is published
	Publisher	**<publish>**	The publishing organization for the dataset
1.2 Description		<descript>	
	Abstract	**<abstract>**	Longer qualitative description of the dataset explaining what is being modeled, the locations (state, county, city, etc.) being presented, the time period being represented, and any other pertinent information or background about the dataset.
	Purpose	**<purpose>**	Information about why the dataset was created, its uses and any limitations.
1.3 Time Period of Content 1.3.9 Time Period Information 1.3.9.1 Single Date/Time		<timeperd> <timeinfo> <sngdate>	Provides the relevant time period for which the data modeling. Can either be a specific date or date range
	Calendar Date	**<caldate>**	Specific relevant date for dataset
1.3.9.3 Range of Dates/Times		<rngdates>	Range of relevant dates for the dataset
	Beginning Date	**<begdate>**	
	Ending Date	**<enddate>**	
1.4 Status		<status>	
	Maintenance and Update Frequency	**<update>**	Frequency dataset is updated *e.g. As needed, Annually, Based on census*
1.5 Spatial Data Organization Information		<spdom>	
*	**West, East, North, South Bounding Coordinates**	**<westbc>, <eastbc>, <northbc>, <southbc>**	The X,Y locations for the four edge corners of the dataset.
1.6 Keywords		<keyword>	
1.6.1 Theme		<theme>	Identifies broad subject / keyword terms describing dataset. Use a thesaurus where applicable for controlled vocabulary. **Repeatable field:** The metadata can include multiple themes, each theme comprised of one theme keyword thesaurus with one or more related theme keywords.
	Theme Keyword Thesaurus	**<themekt>**	**Recommended Thesaurus for ISO Topic Categories:** ***e.g. ISO 19115 Topic Category*** *e.g. None* Use *None* when identifying keywords and no thesaurus.
	Theme Keyword	**<themekey>**	Specific Keywords describing the dataset. ISO 19115 examples include: *boundary, biota, structure, etc.* Free form keywords when no thesaurus is designated *e.g. Congressional, districts, geology*

Checklist: CSDGM GIS Metadata Fields for Preservation

	Friendly Field Name	XML Tag	Description / Example Value
1.6.2 Place		<place>	Describes geographic scope of the dataset. **Repeatable field:** The metadata can include multiple places, each place comprised of one place keyword thesaurus with one or more related place keywords.
	Place Keyword Thesaurus	**<placekt>**	May designate a specific place thesaurus: *e.g. William S. Powell, The North Carolina GAZETTEER, A Dictionary of Tar Heel Places, (Chapel Hill: University of North Carolina Press), August 1984.* May designate no thesaurus *e.g. None*
	Place Keyword	**<placekey>**	*e.g. North Carolina, Wake County*
	1.7 **Access Constraints**	**<accconst>**	Any restrictions to accessing the dataset. Any legal, statutory or confidentiality restrictions for data sharing and use for the dataset should be listed here.
	1.8 **Use Constraints**	**<useconst>**	Any restrictions or guidance on use of the dataset. May contain disclaimers or recommended dataset citation notation.
1.9 Point of Contact 1.9.10 Contact Information 1.9.10.1 Contact Person		<ptcontac> <cntinfo> <cntperp>	Contact information for data originator/authority.
	Contact Person	**<cntper>**	Name of contact person
	Contact Organization	**<cntorg>**	Name of organization with which contact person is affiliated
1.9 Point of Contact 1.9.10 Contact Information 1.9.10.1 Contact Person		<ptcontac> <cntinfo> <cntperp>	Contact information for data originator/authority.
	Contact Person	**<cntper>**	Name of contact person
	Contact Organization	**<cntorg>**	Name of organization with which contact person is affiliated
	1.9.10.5 Phone Number	**<cntvoice>**	Phone Number
	1.9.10.8 Email Address	**<cntemail>**	Email Address
*	1.13 **Native Data Set Environment**	**<native>**	Describes platform, operating system, and version of GIS software used to create dataset. *e.g. Microsoft Windows XP Version 5.1 (Build 2600) Service Pack 3; ESRI ArcCatalog 9.3.0.1770*
**	1.X **Native dataset format**	**<natvform>**	Describes geospatial data format: **Metadata field supplied by ESRI *e.g. Shapefile*
2. Data Quality Information		<dataqual>	Provides historical lineage and source descriptions for the data used in the creation of the dataset. **Repeatable Field.** Citation information
2.5 Lineage 2.5.1 Source Information 2.5.1.1 Source Citation		<lineage> <srcinfo> <srccite>	
	Originator	**<origin>**	Originator (person or organization)
	Publication Date	**<pubdate>**	Publication Date of the Source for the dataset
	Source Title	**<title>**	Title of the Source for the dataset
2.5.1.1.8.8 Publication Information		<pubinfo>	Publication Information of the Source for the dataset
	Publication Place	**<pubplace>**	Publication Place for the Source for the dataset
	Publisher	**<publish>**	Publisher of the Source for the dataset
2.5.1.4 Source Time Period of Content 2.5.1.4.9 Range of Dates/Times		<srctime> <rngdates>	Time Period for the Source for the dataset Range of Dates for the Source
	Beginning Date	**<begdate>**	
	Ending Date	**<enddate>**	

Checklist: CSDGM GIS Metadata Fields for Preservation

	Friendly Field Name	XML Tag	Description / Example Value
	2.5.1.4.1.Source Currentness	**<srccurr>**	The currentness of the Source used to create the dataset
	2.5.1.6 Source Contribution	**<srccontr>**	Information Source for the dataset
2.5.2 Process Step		<procstep>	Describes the processes performed to create the dataset. **Repeatable field** **NOTE:** The Archives' staff member who processes and ingests the dataset into the Archives Repository may add a process step to the GIS metadata to record and document the data transfer and ingest.
	Process Description	<procdesc>	Describe the process
	Process Date	<procdate>	The date of the process was completed
2.5.2.6 Process Contact Information 2.5.2.6.1 Contact Person		<proccont> <cntperp>	Contact information for the person who performed the process <Applic for GIS use? / N/A for preservation>
	Contact Organization	**<cntorg>**	Name of organization with which contact person is affiliated
4. Spatial Reference		<spref>	Describes the map coordinate system **OR** projection. This is important for map display and data interaction and may be unique to each state depending on if using Geographic, Planar, or Locally defined coordinates.
4.1 Horizontal Coordinate System Definition 4.1.2 Planar		<horizsys> <planar>	
4.1.2.1 Map projection		<mapproj>	Will use the Map Projection System
*	**Map projection name**	**<mapprojn>**	Map projection name *e.g. Lambert Conformal Conic*
OR			
4.1.2.2 Grid Coordinate System		<gridsys>	Will use the Grid Coordinate system
*	**Grid coordinate System Name**	**<gridsysn>**	Grid Coordinate System Name *e.g. State Plane Coordinate System 1983*
4.1.2.4 Planar Coordinate Information		<planci>	
*	**Planar Distance Units**	**<plandu>**	Unit of measure
4.1.4 Geodetic Model		<geodetic>	Geodetic Model
*	**Horizontal Datum Name**	**<horizdn>**	Horizontal Datum Name *e.g. North American Datum of 1983*
*	**Ellipsoid Name**	**<ellips>**	Ellipsoid Name *e.g. Geodetic Reference System 80*
*	**Semi-major Axis**	**<semiaxis>**	Semi-major axis
*	**Denominator of Flattening Ratio**	**<denflat>**	Denominator of flattening ratio
5. Entity and Attribute Information		<eainfo>	Details about the information content of the data set, including the entity types and associated data attributes
5.1.1 Entity Type		<enttyp>	
*	**Entity Type Label**	**<enttypl>**	Label for entity *e.g. 1992_NC_Congress_Districts*
5.1.2 Attribute		<attr>	**Repeatable fields** Description of each data attribute associated with the entity
*	**Attribute Label**	**<attrlabl>**	Name of attribute as it appears in the attribute table
	Attribute Definition	**<attrdef>**	Description of the information that is being captured in the attribute field Because the attribute name term may be an abbreviation or mean different things to different people, you should add a description of the attribute and provide the source for that description.
7. Metadata Reference Information			
*	**Metadata Standard Name**	**<metstdn>**	Name of metadata standard used to document the data set.
*	**Metadata Standard Version**	**<metstdv>**	Identification of the version of the metadata standard used to document the data set.