7-2009

A Data Storage Scheme for Geo-Spatial Data Organization and Metadata Cataloging

Jennifer Fu
Florida International University, Fujen@fiu.edu

A Data Storage Scheme for Geo-spatial Data Organization and Metadata Cataloging

Jennifer Fu, GIS Center, Florida International University

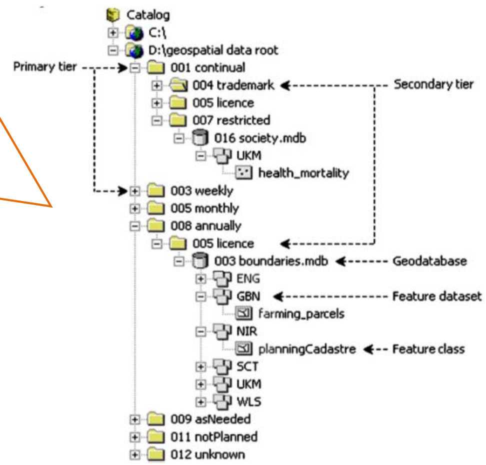
ABSTRACT

This poster explores the possibility of a data storage scheme and uniform knowledge base for geospatial data collections which allows for automated generation of templates for metadata creation and editing.

CURRENT STATUS

Most GIS catalog portals utilize Z 39.50 based distribution search systems, in which standardized metadata cataloging may or may not be followed, quality control is conducted locally, variations in cataloging are common, and there is a lack of uniform data file organization and naming convention. Two practices that metadata cataloging behind the GIS portals misses and that the traditional library catalogs follow are 1) union catalogs for libraries to share cataloging records, and 2) a uniform classification scheme for naming convention, e.g. Library of Congress Classification Scheme, to physically label and organize the materials.

Batcheller (2008) proposed a 3-tiered ESRI geodatabase file structure to uniformly group GIS datasets by their frequency of updates in the first tier, ISO subjects in the second tier, and individual file names in the third tier for the purpose of automated metadata . The major drawback of this approach is its limitation to ESRI's geodatabase platform, and the exclusion of heterogeneous datasets.



The Library of Congress classification is an alpha-numeric code that describes a piece of intellectual property in an abbreviated, meaningful form. The subject/thematic, spatial, and temporal elements are in the sequence that groups materials such as books and journals by its main subject first, then by geography, sub-geography, and last by time. The principle of assigning a linear alpha-numeric code which depicts all important discovery level metadata elements can be adopted for GIS data files for the purposes of not only uniform file naming convention but also automated metadata extraction.

An example is the *Urban Atlas, Tract Data for Standard Metropolitan Statistical Areas: Harford, Connecticut from the U.S. Bureau of the Census, 1974*. Here the discovery level metadata elements such as title, publisher, format, spatial, temporal coverage, and theme are abbreviated into this Call number: *G1242.H3E25U5 1974*

G Geography. Maps. Anthropology. Recreation
1242 Connecticut Sub-area atlas
.H3 Hartford Metropolitan area
E 25 Statistical areas. Census tracts
U 5 Cutter for United States
1974 Date of atlas publication

APPROACH

A sample of geo-spatial metadata codes with some of them based upon ISO proposed elements (UK-GEMINI, 2004) for discovery level metadata elements such as file format, subject class, access, spatial coverage, reference, and scale, are illustrated in the table:

code	Format	Topic	Coverage*	Reference*	Scale*	Access
001	vector	farming	World	Great Britain Grid	<1:5,000	copyright
002	grid	biota	North America	Irish Grid	<1:24,000	patent
003	textTable	boundaries	South America	Irish TM	<1:50,000	patentPending
004	tin	climatology	Asia & Pacific	WGS84	<1:75,000	trademark
005	stereoModel	economy	Africa	…	<1:100,000	license
006	video	elevation	…	…	<1:250,000	Intellectual PropertyRights
…	others	environment	…	…	<1:1M	restricted

ISO-codes and corresponding values — *note: Fields with'*' are not defined by ISO*

The code "001" stands for *vector* in the category of *Format*, for *farming* in *Topic, world* in *Coverage,* and so on. An example of applying this scheme to the name of a spatial dataset that contains vector lines of boundaries of countries of the world published on 02/07/1998

This linear depiction of a data file contains seven basic metadata elements, which can all be automatically expanded during the metadata population process via a look-up table. The advantage of this linear naming structure is that no hierarchical folder structure, as used in (Batcheller, 2008), is necessary since the file name itself contains all discovery level metadata elements. Further, this approach can be applied to other platforms than a geo-database file structure. In addition, a linear code can be used to locate a best matched template within a local or shared knowledge base for further metadata cataloging. This process could include: 1) match assigned code of a dataset against a matrix of codes and their associated descriptive content to generate a partially populated metadata template; 2) use the template for metadata editing; 3) final quality control/validation of metadata xml and 4) upload the final metadata xml for publishing locally or into a consortium of geo-spatial data portal union catalog, which can be served as a uniform knowledge base for peer sharing.

CONCLUSION

The proposed classification scheme which requires more collective efforts to compile and develop, will be largely based upon existing ISO codes, such as ISO 19115 , ISO 3166, and ISO 8601. It cannot only benefit the consistent GIS data file organization by grouping geo-spatial datasets of similar nature next to each other, but also facilitate metadata creation by picking the best matched template from a shared uniform knowledge base. By enforcing standardized metadata creation starting from file naming convention, this could also avoid inconsistent file storage practice within an organization and reduce errors in metadata creation, as the amount of manual editing and original metadata creation is reduced.

REFERENCES

Batcheller, J. K. (2008). Automating geospatial metadata generation—An integrated data management and documentation approach. *Computers & Geosciences*, 34, 387–398.

UK-GEMINI. (2004). UK GEMINI Standard: A Geo-Spatial Metadata Interoperability Initiative (Version 1.0 ed.). London: CabinetOffice, e-Government Unit

FLORIDA INTERNATIONAL UNIVERSITY

AGIT / GI_Forum'09, July 6 -9, 2009, Salzburg, Austria