Assessing the Utility of Current Format Registry Efforts for Geospatial Formats

Nancy Hoebelheinrich; Stanford University Libraries; Stanford, CA/USA & Natalie K. Munn; Content Innovations, LLC; San Francisco, CA/USA

Abstract

Implicit within the metadata strategy of most archiving or preservation institutions is the use of format registries to contain important information, usually technical in nature, that is common among like formats or data types.

At present, there are several approaches to data models, policies, and implementation models for format registries that are in various stages of implementation and/or conception including The National Archives' (UK) PRONOM Technical Registry, the Global Digital Format Registry (GDFR), funded by the Mellon Foundation and led by Harvard University Library, and the Library of Congress (USA).

The National Geospatial Digital Archive (NGDA) project funded by the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP) has been investigating whether existing or planned format registry efforts do or would support the often quite complex geospatial data formats which NGDA and other institutions are collecting for long term preservation.

This paper discusses results of a comparative study of the data models of pertinent format registries in which instances of over 20 proprietary and open geospatial formats were examined to assess whether the elements within the data models could adequately describe a given format and its relatives, and if not, what other kinds of information would be important to include.

The paper places the findings and recommendations into the context of previous work done by the NGDA team and others about what preservation metadata would be appropriate for geospatial resources. In addition, the paper discusses differences in definitions for key format registry concepts describing relationships among formats.

Finally, we identify related research questions yet to be answered regarding the usability of existing and potential format registry efforts for geospatial resources, and the broader questions about the practicability of gathering this and other pertinent preservation metadata for geospatial resources.

Introduction and Overview

The National Geospatial Digital Archive (NGDA) is a collecting network for the archiving of geospatial images and data supported by the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP). As one of the principle nodes in the network, and partners in the NDIIPP funded project, members of the NGDA team at the Stanford University Libraries have engaged in research to identify the information that is considered important to gather in order to archive geospatial data over time. Initial research resulted in a paper documenting the results of an investigation into the need for preservation metadata for geospatial resources [1] The initial investigation found an assumption that the use of format registries were an implicit and important part of the metadata strategy for most archiving and preservation institutions. Yet, from a cursory review, it was unclear how comprehensively geospatial data could be documented within burgeoning data format registry efforts in the US and the UK, and thus the NGDA team decided to build a wiki-based format registry as a temporary measure until research could be done on the treatment of geospatial data in select data format registry efforts. Content Innovations, LLC & Geodata Analytics LLC were contracted to conduct this research under the direction of Stanford NGDA staff.

We researched treatment of 23 geospatial data formats & 13 format subtypes in key format registries and registry related efforts such as PRONOM [2], the Global Digital Format Registry (GDFR) [3] and the Library of Congress' sustainability factors planning matrix [4]. We analyzed sample data targeted for ingest into NGDA and examined how likely the target formats were represented in these key registries.

We also compared format registry data models and mapped common fields and features across the registry efforts of NGDA, PRONOM, GDFR, and the Library of Congress. This effort will aid NGDA either in finalizing its data model for the NGDA format registry, or in deciding which format registry is best suited for documenting the geospatial data that is being archived in the NGDA.

Research: Methodology, data collected, and description of sources

Methodology

For each of the formats in the format research set we documented its full name, short name (if any) file extension convention (if any), and provided a vernacular description of the spatial data format. We prepared notes about application of the format generally as well as locally regarding use of the format in sample data sets at Stanford in California and the West where appropriate.

For each of the formats in the format research set we determined whether the format was a file format or spatial data container. If a container (rather than a file format type) we noted what the container's minimal file components were if known and any additional container components with their likely file extensions.

For each format examined we identified Uniform Resource Locators for the format's de-facto home page, published specification if public, and select whitepaper(s). We noted the version of the specification and/or format where known/germane.

Format registries, while sometimes accommodating containers, are more typically geared toward defining data types at the file format level. Geospatial data, however, is frequently comprised of data sets which may contain multiple files in several different file formats. Thus, discriminating between containers and container elements, and isolating container element file formats were important first steps in our research.

Our research was informed by the abstract notion of a container as "a class, a data structure, or an abstract data type (ADT) whose instances are collections of other objects. They are used to store objects in an organized way following specific access rules". [5]

Our research was also informed by a GIS Specific definition of data format as: "a specific, possibly proprietary, set of data structures within a software system." rather than as a file format specific definition, hence we were able to examine both geospatial datasets as containers as well as specific file formats (often as container elements). [6]

Of course, we also recognized the importance of properly documenting data formats at the file level (e.g. TIFF,) where a file format is defined as: "a particular way to encode information for storage in a computer file". [7] This is a more strict definition of a data format at the file format level. Common methods for identifying file formats include:

- filename extension (e.g. .doc, .xls, .ppt, etc.)
- internal & external signatures
- magic number
- explicit metadata
- Mac OS type-codes (superseded by Mac OSX Uniform Type Identifiers (UTI)
- OS/2 extended attributes (".TYPE")
- MIME types

We recorded file name extensions, linked to specifications where signature information was documented, and noted where explicit metadata existed in container element files or headers. We also noted type and MIME type data fields in our format registry model field mapping effort.

Geospatial Formats Researched

We researched format information for the following spatial data formats:

- Band Interleaved by Line, Component File (BIL)
- Band Interleaved by Pixel, Component File (BIP)
- BLW ESRI Arc View World file for BIL
- Digital Elevation Model (DEM) (in ESRI GRID format)
- Digital Orthophoto Quadrangle (DOQ) in native DOQ or as Geotiff
- Digital Raster Graphic (DRG) TIFF (6)
- ESRI ArcInfo Interchange File
- ESRI Arc/View ShapeFile
- ESRI ArcInfo Coverage
- ESRI Geodatabase
    - ESRI Geodatabase (ArcSDE)
    - ESRI Geodatabase (File-based)
    - ESRI Geodatabase (MDB)
    - ESRI Geodatabase (XML)
- ESRI/GRID
- Hierarchical Data Format HDF (5)
    - HDF EOS Hierarchical Data Format-Earth Observing System
- Landsat
    - Landsat 4 /5: Geotiff
    - Landsat TM: Geometrically corrected NDF product (BIL) aka Landsat 4 /5 BIL
    - Landsat TM: Geometrically corrected NDF product (BSQ) aka Landsat 4/5 BSQ
    - Landsat 7 ETM+ off gap-filled products: Geotiff
    - Landsat 7 ETM+ SLC-on mode: Geotiff
- MrSid Multi-resolution Seamless Image Database
- National Aerial Photography Program (NAPP) in ESRI GRID format
- National Elevation Dataset NED in ESRI GRID format
- Navigational Charts as ARC Digitized Raster Graphics (ADRG)
    - JNC as ADRG
    - ONC Operational Navigational Chart as ADRG
    - TPC as ADRG
- Shuttle Radar Topo Mission (SRTM) as TIFF
- TIFF (6)
- GeoTIFF
- SDTS Spatial Data Transfer Standard
    - SDTS-TVP Topological Vector Profile
- Vector Product Format (VPF)
    - World Vector Shoreline Plus (ESRI Shapefiles or VPF)

Sources & Web Resources

For each format we examined select format registry efforts (NGDA, PRONOM, GDFR & LOC) to determine if the format was defined in the registry. For each format we also looked for treatment, description or examination at the format level by JHOVE, the JSTOR –Harvard's Object Validation Environment, and the Federal Geographic Data Committee (FGDC) and Open Geospatial Consortium (OGC) websites.

For each format we also checked four commonly used GIS conversion tools/utilities (ESRI, GDAL, Manifold, & SAFE) to determine whether the format was supported for import & export and/or direct read & direct write.

A bibliography of our research targets and sources is included in the Full Report in ***Appendix A: Ngda Format Registry Research Bibliography And Resources*** [8].

Format Registry Model Research

To ascertain the comprehensiveness and utility of our own efforts to create a format registry for geospatial resources, we compared format registry data models and mapped common fields and features across the registry efforts of PRONOM, GDFR, the

Library of Congress and the NGDA. We examined the published data models & data dictionaries for these registries (where available).

We recorded common fields used by all or most of the registries we examined and mapped those fields to one another where possible. NGDA's format registry wiki fields were similarly mapped. Where the NGDA registry effort had not yet modeled a particular field, we noted the discrepancy. In a head to head comparison PRONOM and GDFR were most similar, with LOC and NGDA sharing some fields in common. For instance, PRONOM and GDFR both recorded format specific internal and external file signatures and compression while PRONOM, GDFR, and LOC each supported the description of intellectual property rights attached to a given format.

The extensive use of containers in Geospatial data formats requires a format registry that allows descriptions of complex relationships between containers, container elements, related data formats, and format versions. We found that each examined model had some structure to accommodate Relationship to other formats, including

- *Has subtype*
- *Subtype of*
- *Contains*
- *May contain*
- *Used by*
- *Based on*
- *Defined via*

In addition, each examined registry model has some structure to accommodate relationship to other ***versions*** of a given format, including.,

- *Has earlier version*
- *Has later version*
- *Has version*
- *Version of*

We examined the following format registry fields across NGDA, PRONOM, GDFR, & LOC:

1. System ID | Internal Unique Identifier|
2. External Identifier
3. Name
4. Version
5. Alias
6. Family
7. Format Type aka Classification
8. Description
9. Filename Extension
10. Assessment
11. Orientation
12. Byte Order
13. Grammar
14. Related File Formats
15. Internal Signature
16. External Signature
17. File type signifiers
18. Compression Type
19. Character Encoding
20. Format Disclosure
21. Release Date
22. Withdrawn Date
23. Developer aka Developed By, Created By
24. Support aka Supported By, Maintained by
25. Documentation
26. IPR
27. Caveats
28. Notes: General
29. Notes: History
30. Reference File
31. Local use
32. Production phase
33. Transparency
34. Self-documentation
35. External dependencies
36. Technical protection considerations
37. Internet Media Type
38. File type signifiers

Findings

The Geospatial formats we researched were well represented in popular industry software and conversion utility packages. More than 50% of the formats we examined were accommodated types in GDAL, Manifold & SAFE's conversion utilities. At least 70% of the formats we examined were directly read/writable by ESRI software or accommodated by ESRI's Interoperability Extension.

Registry efforts at PRONOM and GDFR are setting the standard for modeling and publishing data format definitions, but these particular registries have impoverished and incomplete coverage of geospatial data formats at present (although we looked at the GDFR registry at a very early stage when not all of the data had been included in the public view of the registry). Less than 1/3 of the data formats we examined were present in PRONOM while LOC & GDFR had even fewer – as few as two or three formats represented in each. The more complete survey results of our comparison can be found in ***Appendix B: NGDA Registry Survey***. [9].

In addition, the complete Registry Field Map research findings are presented *in* ***Appendix C: NGDA Registry Field Map Research.*** [10].

Our research found that both PRONOM and GDFR's registry data models handle a high enough level of abstraction to accommodate the challenges outlined above. While both data models are valid, it was more difficult to understand the full GDFR data model because GDFR's naming conventions and level of abstraction combine to make it more difficult to decipher their registry entries out of context of the larger GDFR data structure. From feedback received from GDFR developers, we have since learned more about some assumptions underlying the GDFR data model, thus giving us a better understanding of that data model.

As an informative exercise, we prepared format definitions for two formats using the PRONOM model. For the purposes of communicating with our audience, presenting example registry definitions in XML against PRONOM's model made them more easily human readable. See ***Appendix D: Sample Geospatial Format Registry Definitions*** [11] for examples of draft format registry definitions for select geospatial formats.

We found that taking a quick look at the same format defined side by side in each registry was one way to get acquainted with each registry's XML output and get a feel for the way the underlying data structures influence the definition level entries in GDFR and PRONOM. See ***Appendix E: GDFR and PRONOM Format Registry Definitions' comparison*** for a side-by-side comparison of TIFF format registry definition in XML for PRONOM and GDFR's registry data models. **[12]**

None of the registry efforts examined support links either to archived copies of referenced specification or white papers, or to sample files for each file format entry. The NGDA team thought that would be a very important function of a format registry so that those interested in looking at source information could most easily find it.

Report Recommendations

As previously noted, it is particularly important for geospatial data that format registry efforts adopt data models for their registries that accommodate parent child relationships between containers and container elements, as well as relationships between format versions and related data types. As well, support for automated ingest into an archive with a commensurate data format validation step in the ingestion process would demand that geospatial data format definitions be authored down to the container element and component file format level. For an automated process to work well, all likely container element and file types would be accurately described in the registry for any given registry entry. NGDA's draft format registry data structure would need to be revised to fully support containers and parent/child relationships.

Both GDFR and PRONOM's data models are valid and either can accommodate data format registry entries for geospatial data types; yet, geospatial data are not well represented in these registries. NGDA is debating whether to adopted a revised data registry model similar to GDFR's, or implement a mirror node of GDFR to populate and test geospatial format registry entries against their model.

A feature that is considered important by the NGDA team is the capability for populating a format registry so that it supports and includes links to reference/sample files for each file format entry.

One feature that we found useful was GDFR's "save to XML" which appears to write tags for populated attributes from GDFR's base, format, and product tables. Less useful was the fact that if no value was entered for a particular registry attribute in GDFR, the empty tags for that attribute don't appear in GDFR's "save to XML". Since GDFR doesn't output empty tags, the default export/display of GDFR's minimal level entries in XML does not allow the casual user an immediate way to determine which elements of a given registry definition have not yet been completed.

Were the NGDA team to expand its current wiki based format registry, we would implement an "export to XML feature" that would support easy self- export & registry entry portability. An export to XML feature that allows the user to select whether to write empty tags for undefined attributes would be a useful registry enhancement.

The NGDA team is investigating the feasibility of providing the enhanced or unique geospatial data format definitions that are being compiled during the research phase of the format registry investigation to LOC, GDFR, and PRONOM's registries as an output of this project concurrent with or instead of their publication in an NGDA wiki/format registry. As noted, few geospatial format types are actually populated at the moment in any of the format registries examined, and mechanisms for adding format definitions are not at all clear or easy to use at the moment. One of the reasons that the initial NGDA format registry was wiki-based was to encourage the geospatial community to contribute format definitions despite the lack of clear authority that is inherent in a wiki based mechanism. Further developments will reveal whether the authority based or community based mechanisms for authoring format registry definitions will be the most feasible.

Additional Research Questions

With a few moderations, it appears that data models of existing or developing format registries would prove suitable for recording registry definitions of geospatial formats. What is not readily apparent, however, is how the format definitions will be populated.

One aspect of the continuing research of the NGDA project will be the evaluation of the feasibility of authoring and contributing format registry definitions for the formats ingested into the NGDA. As part of the evaluation, the team will continue its investigation of how the data models work for geospatial resources, particularly with regard to the use of container and file component elements. In addition, the team will evaluate how practicable it is to collect and archive format specifications, white papers, and instances of geospatial formats in the public domain that can be referenced as samples for the format registry definitions. It will be particularly interesting to see how proprietary formats can be documented as these are quite common in geospatial resources, and it is not clear that a general knowledge or appreciation that such information is important for long term preservation of geospatial resources within the geospatial domain.

There has been some discussion of late in various venues about whether the kind of information found in format registries is useful or necessary for long term preservation. [13] As that discussion continues, some practical experience in authoring and contributing format definitions should provide a useful perspective on the practicability of this kind of work. In addition, the NGDA team plans to provide some metrics on the presence of preservation metadata for geospatial resources as it continues to ingest such materials into the NGDA.

References

[1] N. Hoebelheinrich, et al., "An Investigation into Metadata for Long-Lived Geospatial Data Formats", www.digitalpreservation.gov/news/events/ndiipp_meetings/ndiipp08/docs/session7_hoebelheinrich_paper.doc. (2008).

[2] PRONOM. The Technical Registry. http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

[3] Global Digital Format Registry. http://www.gdfr.info/

[4] Sustainability of Digital Formats: Planning for Library of Congress Collections. http://www.digitalpreservation.gov/formats/

[5] "Survey And Assessment Of Sources Of Information On File Formats And Software Documentation Final Report" The Representation and Rendering Project University of Leeds. http://www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf (May 20, 2003).

[6] "Annex VI. Glossary" from "Handbook on Geographic Information Systems and Digital Mapping", Studies in Methods, Series F, No. 79, United Nations Department of Economic and Social Affairs, Statistics Division, New York, 2000,.

[7] File Format. http://en.wikipedia.org/wiki/File_format

[8] Content Innovations, LLC, "Appendix A: NGDA Format Registry Research Bibliography and Resources" from "Report to National Geospatial Digital Archive Regarding Geospatial Data Treatment in Data Format Registry Efforts". http://www.contentinnovations.com/ngda/regfindings.html. (November 20, 2008).

[9] Content Innovations, LLC, "Appendix B: NGDA Registry Survey" from "Report to National Geospatial Digital Archive Regarding Geospatial Data Treatment in Data Format Registry Efforts". http://www.contentinnovations.com/ngda/regfindings.html. (November 20, 2008).

[10] Content Innovations, LLC, "Appendix C: NGDA Registry Field Map Research" from "Report to National Geospatial Digital Archive Regarding Geospatial Data Treatment in Data Format Registry Efforts". http://www.contentinnovations.com/ngda/regfindings.html. (November 20, 2008).

[11] Content Innovations, LLC, "Appendix D: Sample Geospatial Format Registry Definitions" from "Report to National Geospatial Digital Archive Regarding Geospatial Data Treatment in Data Format Registry Efforts". http://www.contentinnovations.com/ngda/regfindings.html. (November 20, 2008).

[12] Content Innovations, LLC, "Appendix E: GDFR and PRONOM Format Registry Definitions' Comparison" from "Report to National Geospatial Digital Archive Regarding Geospatial Data Treatment in Data Format Registry Efforts". http://www.contentinnovations.com/ngda/regfindings.html. (November 20, 2008).

[13] David Rosenthal blog post on format specification. http://blog.dshr.org/2009/01/are-format-specifications-important-for.html
and Chris Ruthbridge response on DCC blog. http://digitalcuration.blogspot.com/2009/01/specifications-again.html

Author Biography

Nancy J. Hoebelheinrich, *Stanford University Libraries is Metadata Coordinator for the Digital Library Systems and Services department at the Stanford University Libraries / Academic Information Resources. In that capacity, Nancy coordinates metadata services for Stanford Libraries' digital production activities, digital repository development and implementation, and educational technology services. She has been a member of the METS Editorial Board since 2002 and is currently serving as co-chair. Nancy has been active in a number of information and educational technology specification efforts including that of PREMIS (for preservation metadata), and several of IMS Global specifications related to packaging, repository and resource list interoperability. She is currently involved with the IEEE Learning Technology Standards Committee's RAMLET project, and continues to monitor various groups working on practices related to the use of digital rights expression languages.*

Natalie K. Munn, M.A., M.L.IS., Content Innovations, LLC Principal, is an Information Systems Specialist with over fifteen years experience. She works with libraries, museums, corporate, and academic clients to implement new systems, and convert databases to more modern systems. Ms.Munn has specialized experience with library catalogs, controlled vocabularies/taxonomies, geographic information systems, multi-media and image databases. Ms. Munn performs system analysis, database design, and database enabled intra/internet projects.