



## Report on Status of Format Registry - 2005-09-19

Our task was initially defined as "Analyze spatial data formats and develop local format registry with programmatic interface to existing format registry". Originally, the first part of this task was headed up by Meredith Williams of Stanford and, the latter part, by Catherine Masi of UCSB. We have since determined that the first part is a research task, rather than a technical task, and has thus been subsumed under the Analysis and Research subgroup of the project, headed up by John Banning. The progress on the latter part of the task is described below:


### Develop local format registry

We determined through research and in conformance with our technical architecture that the high level technical requirements of NGDA Format Registry were that it would:

- Be independent and self-contained within the archive
- Contain sufficient semantic information to programmatically access format
- Contain definitions which would exist in simple documented file formats in a simple directory structure
- Have an access/search mechanism (GUI/database) built above it but this would not be necessary for access to basic definition files

Eventually interface with collaborative authoritative FR for updates and contributions

We studied what format registries currently exist. The most relevant that we found were the following:

- Library of Congress Digital Formats 
- Global Digital Format Registry (GDFR) - Harvard 
  - Global Digital Format Registry Description
  - Ockerbloom's Format Registry Demonstrator (FRED) 
- PRONOM - File format registry - UK archives 

Practical, in use, not geo-spatial

We studied if the existing format registries contain geospatial formats and found that they do not. Our task is to add geospatial formats to an existing registry effort such as the LCDF. We then decided to focus preliminarily on the LCDF since our grant is to archive data for the Library of Congress. We studied if LCDF supports access and contribution mechanisms, how the formats are stored internally and whether a data dictionary exists to define the fields. We determined that there are no access and contribution mechanisms at

this time, that the formats are stored in MS Word files and the data dictionary is [http://www.digitalpreservation.gov/formats/fdd/fdd\\_explanation.shtml](http://www.digitalpreservation.gov/formats/fdd/fdd_explanation.shtml).

Our next step was to coordinate our efforts with the LCDF, GDFR, FRED, TOM. Since Nancy Hoebelheinrich of Stanford was planning to attend the DLF meeting, she initiated contact with Stephen Abrams of Harvard (GDFR), John Ockerbloom of Penn (GDFR, FRED) and Steve Morris of NCSU and met with them at DLF. We came up with the following questions regarding the technical aspects of building a local format registry and possible interaction mechanisms between LC, GDFR and our local format registry:

- What vetting is in place to establish the authority of entries in the LCDF?
- How are entries maintained? Who is responsible for submitting corrections, changes to versions, etc.?
- What is the relationship of the LCDF to the GDFR and University of Pennsylvania efforts (TOM & FRED)?
- Where and how are the actual format definitions stored? How might an interactive (service) be set up that would provide automated communication between local repositories and the LCDF?

Unfortunately these questions were not addressed very well at this meeting. Catherine Masi subsequently contacted Stephen Abrams (Harvard - GDFR) and John Mark Ockerbloom (Penn - FRED), to open up a discussion on the technical aspects of developing a geospatial format registry. S. Abrams responded that the GDFR is still only an idea rather than a reality and that a technical discussion of how our GIS formats should be managed in a GDFR-conformant way is a bit premature.

Thus the task has now become:

**"Develop a local format registry focusing on technical specifications and validation information with a view toward building a community of experts to contribute formats and definitions in the future and with a view toward interfacing with a collaborative format registry such as GDFR when it is implemented".**

The UCSB Technical Architecture team met continuously and discussed the physical structure and content of the format registry. Catherine began prototyping the physical structure using the CASIL (<http://gis.ca.gov/data.epi>) formats. During the course of building the registry we discussed, analyzed and completed work on the following issues:

**Physical structure of registry:** Catherine created a hierarchical directory based registry and included the entire format spec locally, (e.g. as a website in the case of geotiff, as local pdf file in the case of shapefile). We subsequently decided to flatten the hierarchy of the directory structure because tfw, for example, is not a subtype of geotiff but can be attached to a tiff or another format.

**Format record layout:** We decided that the most significant information to be included in the registry is the format spec but that validation information could also be useful. We created the record layout with the following fields for our prototype with the idea of adding fields as necessary: format\_name, description, related\_formats, specification, supporting\_documentation, validation\_code, validation\_documentation, file\_extensions.

**Completeness/independence:** All links on the format record refer to local copies of format information. All documentation about the format is located locally in that format's directory. Catherine continues to work on making sure that the format specs are complete and all information

is located locally where possible. Several specs have several layers of documentation so this task has not been trivial.

**Format of format record:** We discussed the format of the basic record and decided that xml with a simple xsl stylesheet would be most appropriate for sustainability.

**Supporting formats:** We decided to include supporting formats (such as html, gif, gzip, css) as well as geospatial formats in our format registry for sake of completeness.

**Format of spec itself:** We decided that we must make every effort to make the format spec itself readable in 100 years and that the best way to do that is to "dessicate" any complex format down to either plain text or a very simple raster image (e.g. gif). We decided that a format spec in a more complex format such as html or pdf was acceptable as long as that format was then defined in either text or gif format.

**Validation:** We are using JHOVE (<http://hul.harvard.edu/jhove>) to validate formats where possible. We have implemented JHOVE locally and included the JHOVE validation code in the format registry as well as the validation module documentation. We discovered that JHOVE's audit tool may be useful for initial identification of format as well so we may it as part of the ingest process at a later date.

**Format records are currently available for bmp, css, geotiff, gif, gzip, html, jpg, mrsid, pdf, png, shapefile, svg, tfw, tiff, xhtml, xml and zip formats. Validation is also available for geotiff, jpg, pdf, tiff, xml. We are concentrating on technical specs and validation only; we will add further description and use notes later.**

**Next steps:**

- All format registry files must reside locally (look at png).
- Do we have ALL the CASIL formats - use JHOVE to determine.
- Review and polish entire format registry.
- Make available in ARCHIVAS and via GUI (create manifest.xml, create update tool, etc.) - assigned to Chris Barteau
- Work on how to validate formats not covered by JHOVE, in particular shapefiles
  - Extend jhove to validate shapefile, possibly after zip - assigned to Justin Mathena
  - Follow up with Safe Software (any opportunity for collaboration?)
- Look into "in vivo preservation" Mike Nelson, Old Dominion (from NDIIPP meeting July 12-13) as way to update format specs.

Research the SourceForge model and other models for building a community of experts to create a robust Format Registry