

## **Global Digital Format Registry (GDFR) Meeting Minutes**

Archives I, Jefferson Conference Room

10 July 2008

2:00 p.m. – 6:00 p.m.

### Introduction

Meeting opened by Robert (Bob) Chaddock with individual participant introduction.

### Purpose of Meeting

Dale Flecker reviewed purpose of meeting and stated that in order for this project to 'live' enough community participation and concern for a format registry is needed. Money is still left in the budget to cover, if needed, the expenses of participants. All expense statements must be received before the end of July. The GDFR effort is at a critical point. The Mellon funding was used on the technical aspects of the Registry, namely to develop a software platform for the Registry. Someone among the GDFR participants needs to throw their energy behind this effort and test the work. It was also mentioned that the National Library in Canada is hosting a meeting with other National Libraries; NARA volunteered to host meeting of various national archive [participants]. In summary, the purpose of this meeting is to test if GDFR is a community and to discuss whether other entities should be invited to join as well as to discuss the ongoing pilot.

### Roles

Harvard and NARA discussed how they perceive their roles...

Harvard: (Flecker) Participating in this project as being an agent for the community; will run for ~ 2 (two) years as support; funding for project runs out and the end of July [2008]; all other research done is at Harvard's expense; No additional funding is expected until GDFR matures; Harvard will serve as pilot project's 'home'.

Aim of Harvard is as follows: (1) Encourage help from community and (2) Help push pilot forward (once it gets a momentum of its own, Harvard will step back and serve the role as a participant). Should any one institution (which one) host?

NARA: (Chaddock) During pilot project, role of Archives will be as a catalyst to support discussion and discernment; Archives interests in serving as an agent for International discussions (International Governance Workshop). Value and opportunity to de-couple activity of pilot; governance involves government; NARA will agent with International Institutions; Decoupling includes: (1) University and (2) Multi-national government's opportunity for mutually beneficial complimentary role.

### Background/history

Stephen Abrams provided a background to project. The Internet Assigned Numbers Authority (IANA) Multipurpose Internet Mail Extensions (MIME). Participants were uncomfortable with the effort being controlled by any single nation.

### GDFR Governance Workshop

Richard Steinbacher advised that all data has been downloaded to the ERA Research webpage; GDFR must be a self-sustaining entity; NARA researched governance models and that was the exploration topic for the November 2007 meeting that was held to share ideas; link for website

sent out in an email; all documented in proceedings report; will serve as a solid foundation for future activities.

### Architecture

PRONOM vs. 4 will be discussed as basis for GDFR model; in some instances both have been ‘overlapping/leap-frogging’ each other. Need to determine if right issues/documents/files can be managed in GDFR directly; based on the OCLC IWSA/RFA framework, relational database could be plugged in, if all works; all software deliverables are being released under LGPL license. Mr. Abrams’s opinion is that the data model is what is going to survive going forward and the technology itself is ephemeral. The files that instantiate the documents can be managed within the GDFR space, but more likely, will just need to be pointed to.

- Interoperable Services Architecture
- The Framework Architecture
- Public Service Layer supports any number of TCP-based applications.
- So far worked with Object Oriented Database, but could work with a Relational DB as well.

### Current state

At end of this month [July], OCLC is wrapping up final programming and turning over software to Harvard for maintenance. Software will transition from OCSC to new GDFR website; information about GDFR will be separated (current vs. future); nodes are located on GDFR. info website. Two (2) interfaces exist: machine and human/user. Software has been populated with test data (Magic database); will take about a half-day to install/configure; mirror nodes can not currently have local data. OCLC still working on website for remainder of [this] month [website was displayed to meeting participants].

Comments mentioned during review of website included, but were not limited to:

- ~Examples are needed for fields that are fully populated
- ~Mirror node will show/display host information
- ~Local host serves as an identifier
- ~Duplicate records will be discussed later; currently not able to differentiate duplicate records
- ~Able to go directly to record via URL
- ~Concern expressed that OCLC will work on system (development) until very last day; no one is currently familiar with coding of software; concerned development will get ‘handed off’ and no one will be familiar with it
  - ~No one will be ready to use software prior to August 1<sup>st</sup>
  - ~What happens when we change it? Anything that is added is completely unreliable at this point (would have been nice to see a prototype)
  - ~Are we piloting the editorial process or the system itself? → Response: Harvard will be maintaining the database
  - ~Are there enough pieces of the system to make it viable? → Response (Goethals): Part of the pilot is evaluating what is currently available; Pilot includes data & editorial piece
  - ~Software shouldn’t be an issue, but rather the data behind the system

What are we piloting? Did we pilot the governance structure, i.e. an editorial system, or the actual system?

OCLC will drop the code in Harvard’s lap. Harvard will be maintaining it going forward.

Are there enough components to the system?

This is the first time to evaluate and look and to do gap analysis.

Four (4) things to consider:

1. Data Model;
2. Editorial Process;
3. Software;
4. Governance Model.

### Relationship to PRONOM

Technical registry; housed by National Archives of UK

There is no common model with PRONOM.

Questions posed included:

~Do we need two different registries?

~Do we have any sense that PRONOM has an editorial process? Do we have any evidence of a consistent approach?

Are there discordant views between GDFR and PRONOM? We do not know because the consistency [between the two] has never been checked.

If we run a collaboration multi-author effort, we could evolve discordant interpretations until the editorial process cleans it up.

It was stated that GDFR takes a very expansive view of what is a format (anything can be a format); an entire file system can be considered a format if you chose to set it up that way...

It was mentioned that PRONOM stated they have 'bottle-necked'... there is a very strong willingness to work on a relationship between European community (PRONOM) and US community (GDFR)

Project plan for GDFR lists PRONOM populated with GDFR as a milestone

It appears, via website that PRONOM is copyrighted (it was stated that factual data can't be copyrighted; a question was also posed of what would be the UK's motives for copyrighting the software; no one was able to address/answer question at this time

It was also stated that the software (PRONOM) is being commercialized (this statement was flagged as an issue)

National Archives of US will be asked to discuss PRONOM vs. GDFR with National Archives of UK

Sub-typing will be a focus; must agree on format/language/set of names, etc. so that community can 'talk' to each other

Sub-typing in the formats may be the most significant in determining the mapping tables in the tools and facilitating interoperability in the tools.

Multiple anthologies do exist and as long as we could map them, perhaps 85% perfection is just fine.

### Issues and Observations

Question posed: Are we still convinced we need a format registry? → Response: Yes!

Other statements included, but were not limited to the following:

~Participants, as a community, must collectively form 'energy' together to tackle project

~PRONOM appears to be an ‘individual’ project rather than a ‘collective’ one; however, there is a critical assumption that a community will come together to maintain GDFR; a statement was made that having the registry controlled by one entity is not a comfortable situation.

~Mention of investing a lot of effort in the governance portion of the project included a suggestion of getting several pilots working from the ground up (could be beneficial); in addition, an opportunity to see a working prototype would be beneficial as well

~ Fedora community is developing “Solution Councils.” If we form a “Solution Council” in Preservation, we will run into the issue of European Community relying on PRONOM.

~One of the values in community activity is bringing in a whole set of expertise

~Don’t have ‘energies’ to be experts in the formats being used

~Need to collect documents in original form

~I have format ‘X’ and I need format ‘Y’, how do I get there?

~ A metadata person can interview a format expert.

Without governance who is to arbitrate “What is data and what is spam?” Exactly what needs to be done is strongly dependent on the nature of the domain.

### Use cases

Three (3) sets of Use Cases were presented in the handout. Variety of repositories, not just GDFR. A Service Model was also done.

Issues:

1. How evaluative should GDFR be? Should it accommodate assessments – such as the quality of preservation?
2. Should GDFR provide services, such as Notification?
3. The Use Cases are somewhat mechanical.

During discussion, comments included, but were not limited to the following:

~Repetitive question: How add-valuable is GDFR?

~Should these services [use cases] be part of GDFR?

~Need to define scope of GDFR

~What kind of data do we need? How vigorous? How much of an editorial process?

~Perhaps pilot has to be registry, plus evaluation

~Need to know what package produced record

~ DRIOD has demonstrated capability to use file identification tool, but only on a part of the formats.

~Where do you record preservation?

~Want to put in format registry: information that drove the choice

~Business contacts are helping define the format

~Missing the way of integrating tool and using in GDFR

~Format identifier tells what we have; will wait later to identify risk

~Assumption: Make JHOVE 2 ‘speak’ GDFR

~If we have needs that diverge from GDFR then we have to maintain our own local extensions

~Need enough documentation to explain the names

~Can’t use Wikipedia; need ability to upload various files/versions

~Community has more responsibility to be experts for common formats

~Agreed that JHOVE will be able to kick-out GDFR identifiers; at some point community needs to say these identifiers are persistent

- ~Community needs to have a conversation about identifiers; agree that there is a basic gap operation
- ~Topic for next conference call: What these identifiers can be used for...
- ~Have to share database; will need to 'war' through vetting editorial process; Community needs to discuss how they would repository formatted in environment
- ~Challenge will be to get people to put data in database
- ~Need to decide classification: What is a sub-type? What is not?
- ~Database should offer ability to return and re-classify data if need be

### Discussion of pilot

Can Harvard just propose the scope of the Pilot?

1. Enter the data;
2. Do a show;
3. A people-centric pilot, volunteers go away and figure out the scope.
4. A Limited Pilot idea:  
Explore the taxonomic issue of establishing what a "subtype" is. It is often not a trivial question. The need is to differentiate at the ingest time.

Activities 1, 2, 3, and 4 can be performed in parallel. Parallel activities could be solving the tools and editorial issues. Solving the editorial issues would allow reliance on the data.

Suggestion that activity "2" is a Federation Activity (*note takers: not certain if this was recorded correctly*)

It was stated that Community needs to determine whether National Archives (NARA) will be willing to communicate with National Archives of UK regarding PRONOM

Community agreed with purposes one (1) through three (3) for pilot

Is the schema for PRONOM available?

GDFR schema is a script superset of PRONOM4.

There will be a node for every entry in the GDFR repository.

Need to adjust the Harvard repository from using PRONOM as a placeholder and transition to GDFR.

Will the Pilot tool allow editing the content?

When the root node goes live in August 2008, will the attendees get accounts on the tool?

The tool allows editing, but does not allow dialog. Experience can be recorded in a wiki.

Portico (Evan Owens) volunteered to mirror, once the data is stable.

NARA volunteered to explore the governance issue.

Request:

Would NARA consider assigning an officer to communicate with TNA about having a mutual agreement for sharing of PRONOM data?

Dale and Andrea will write up the proposal; will discuss with National Library by mid-August; will have to Team/participants well in advance of meeting [with Nat'l Library]

A target to start in September [2008] was set

It was mentioned that additional thoughts can be emailed to Dale and/or Andrea (Harvard)

---