Proposal to the National Digital Information Infrastructure and Preservation Program (NDIIPP) for Technical Development Support for the Unified Digital Format Registry (UDFR)

Contents:

I. Abstract

The Unified Digital Format Registry (UDFR) will provide a reliable, sustainable and publicly accessible knowledge base of file format information for use primarily by the digital preservation community. While the UDFR project formally began in April 2009, it is based on a decade of format registry work performed by a number of institutions internationally (see *Appendix B*). The UDFR project was created to unite the separate PRONOM and Global Digital Formats Registry (GDFR) projects.

PRONOM, created by The National Archives (TNA) in the UK, has a strong technological base, and has been building a database of original information about various digital formats. PRONOM however is owned and maintained by a single organization, making it vulnerable to changes in that institution and reliant on TNA for all data contribution. The Global Digital Formats Registry (GDFR) effort, hosted by Harvard University, has developed a model for a registry based on shared governance, cooperative data contribution, and distributed data hosting. However, GDFR is technically less far along in development, and has not yet begun production database building.

Recognizing the multiple benefits in combining the PRONOM and GDFR initiatives, a group of involved and interested institutions agreed to join together to create the UDFR, drawing on the individual strengths of the two existing efforts. The UDFR is currently governed by an interim body of interested institutions (see Appendix A) but is expected to transition to a permanent governing body and hosting institution in late 2009 or early 2010.

The work required to establish the UDFR falls in two general categories: governance and technical work. The governance work includes designing and implementing the plan for on-going registry governance, funding, and operations. The technical work includes the design, development and testing of registry software and processes needed to exchange registry information with tools, services and repositories. This proposal concerns solely the UDFR technical work. The governance work will continue in parallel to the technical work described in this proposal.

This proposal lays out a one-year plan of work to be conducted at the California Digital Library (CDL) of the University of California, leading to the operational deployment of the first version of the UDFR registry. It will support the core features of the registry, including the ability to search, browse, add, modify and export registry content. It will support access to registry information from both a web user interface as well as from published web-based APIs, and will be populated with format and environment information from the PRONOM database.

II. Background and work to date

This project will build upon a large body of previous work conducted for the GDFR and PRONOM projects. It will also seek to reuse existing open source software and technologies to minimize the technical effort needed to realize a first version of the UDFR registry.

1. Data models
    a. GDFR data model
        i. The GDFR data model is defined in terms of a number of high-level conceptual entities, such as Format, Agent, Hardware, Media, Software, and Document. It also models inter-entity relationships such as dependencies, authorship/ownership, assessments and documentation. The complete data model is available on the GDFR website[1].

    b. PRONOM 6.2[2] data model
        i. The PRONOM 6.2 data model is similar to the GDFR data model. The main differences are:
            1. In the PRONOM data model encodings and compression types are treated as explicit entities; in the GDFR data model these are treated as formats.
            2. PRONOM does not contain GDFR's Media, Grammar or Assessment entities.
            3. PRONOM's data model supports migration pathways; GDFR's does not.
            4. The GDFR data model contains elements that were intended to support an editorial process; PRONOM's does not.
        ii. The PRONOM 6.2 data model is not publicly available but is available[3] for use by the UDFR project.

    c. UDFR data model
        i. The UDFR data model is currently being finalized by the UDFR Technical Working Group. It is a combination of the elements shared by both the UDFR and GDFR data models, plus additional elements needed to support the UDFR use cases. The elements present in PRONOM but not GDFR, and vice-versa, are being re-evaluated by the group for consideration in the UDFR data model.
        ii. The documents related to the UDFR data model are located on the private UDFR wiki.

---

1 Harvard University Library (May 22, 2008), Data Model, v. 5.0.14, <http://www.gdfr.info/docs/GDFR-data-model-5_0_14.pdf>.

2 PRONOM 6.2 is the latest version of PRONOM for which the IP clearly is owned by the UK National Archives alone. A commercial company, Tessella, retains some IP ownership for PRONOM versions 7 and 8. For this reason the UDFR Interim Governing Body determined that versions of PRONOM later than 6.2 would not be appropriate for use as a base for UDFR.

3 Tessella (January 15, 2007), PRONOM 6 Architectural Design Document, V1.R6.M0, <on the private UDFR wiki>.

2. Software
    a. Semantic wiki software
        i. The concept of semantic wikis has existed for almost a decade, but it has only been in the last few years that they have matured to the point of being considered as platforms for production systems. Semantic wikis have the ease of use that wikis do but differ from wikis in that they allow structuring, annotating and linking text to support machine-based queries and exports of the content. Semantic wikis are useful for housing knowledge bases such as the UDFR registry.
        ii. Use of a semantic wiki as the underlying software for UDFR could have several advantages over using PRONOM 6.2 or the GDFR software:
            1. The relatively simple interface could promote data contribution from a larger population.
            2. Most of them already support standard APIs for query and export of the data. This capability is needed by tools and services to retrieve information from the UDFR registry.
            3. To the extent that customization of existing semantic wiki software is not needed or at least minimized, the UDFR organization can avoid devoting a lot of resources to maintaining and enhancing custom UDFR software. Over time, as needed, the UDFR content can be imported into newer and better versions of semantic wiki software.
        iii. There are a number of available open source and commercial semantic wiki software packages such as Semantic MediaWiki, SMW+, OntoWiki, SweetWiki and IkeWiki. These and other implementations are easily found through web searches.

    b. PRONOM 6.2
        i. PRONOM 6.2 was created by Tessella under contract to the UK National Archives (TNA). TNA singularly owns the IP rights to PRONOM 6.2 and has offered it as the technical base for the UDFR registry. PRONOM 6.2 is full-fledged repository software - the registry portion of it is a relatively small portion embedded within the software. It is written in ASP.NET, Transact SQL and XML. It requires IIS and SQL Server 2005. Administration is performed via an MS Access application.
        ii. In addition to the technical knowledge-base it includes web services to:
            1. export DROID signature files
            2. support characterization decisions (used by a PLANETS characterization service)
            3. support risk-based preservation planning and migration (used by TNA's Seamless Flow program and currently not exposed in the public instance on the PRONOM web site)
        iii. The software can be obtained for the UDFR project by request to TNA.
        iv. Potential advantages of using PRONOM as the base include:
            1. Presumably the PRONOM data that will serve as the software

base for UDFR data is more compatible with the PRONOM software than the other software under consideration.
2. This version of PRONOM has already been in production for a few years so it would require less testing than the other software under consideration.
3. It already supports exports to DROID, which UDFR needs to support

v. Potential disadvantages of using PRONOM as the software base includes:
1. the complexity of the software (because it is embedded within general repository software)
2. it requires commercial software
3. its written in somewhat outdated technologies
4. the data input interfaces would have to be rewritten so that they are web-based

c. GDFR

i. A GDFR implementation was created by OCLC under contract to Harvard University. It is written in Java, XML, XSLT and Perl. It requires:
1. Tomcat 5.5.25
2. Apache 2.0.52 with mod_perl, mod_rewrite, mod_jk
3. Berkeley XMLDB 2.3.10
4. Java 1.5 JDK
5. Perl 5.8.5
6. Apache Ant 1.6.2
7. GCC 3.2 (for compiling XMLDB only)

ii. It provides a web-based interface (requiring the FireFox web browser) supporting search, browse, record display, record addition or modification.

iii. Potential advantages of using the GDFR software as a base is its written in current software languages and the software it depends on is open source.

iv. Potential disadvantages of using the GDFR software as a base includes:
1. it has never been fully tested and the testing that was performed uncovered many bugs in the software - the amount of work needed to get it fully functioning is unknown
2. the data input forms are complicated and require strong knowledge of the underlying data model - this may discourage contribution beyond a limited number of people

v. The software is available for download on the GDFR website[4].

3. Use cases

Extensive work has been conducted as part of the GDFR and UDFR projects to gather format registry use cases from a broad range of institutions. Those use case are

4See http://www.gdfr.info/download.html

available on the GDFR and UDFR web sites and wikis. The UDFR Technical Working Group compiled the use cases and selected the core set to be completed in this project for the first version of UDFR. This functionality is described in the *Program of new work* section of this proposal.

4. Data
   The PRONOM data will be used as the base for the UDFR data. A search in the PRONOM interface shows that it currently documents 615 formats. These can be exported in XML or CSV formats from the PRONOM web interface. Note that the elements that are exported differ between the XML and CSV formats.

III. Program of new work

At the end of this one year project the first version of the UDFR registry will be operational and available for use. The UDFR will be initially deployed as a single instance hosted by the UC Curation Center at the California Digital Library (CDL). However, the long-term goal of the UDFR project is to support automated replication of UDFR content between instances operated by various institutions. (The work necessary to implement the distributed UDFR network is outside the scope of this proposal.) The CDL instance will run in a high-availability cluster with automated failover. Conforming with IT best practices, the primary copy of UDFR content will be stored on RAID disk with nightly backup to tape.

The central registry will support data queries, single record export and batch exports by requesting tools, services and repositories via a web-based user interface and via publicly available APIs. Through the batch export mechanism, institutions will have a method to obtain the registry content either for additional safekeeping of the content, or for use in local applications.

The central registry design will include storage for the registry records, related documentation and specifications, and test files (example files per file format). The registry records will be capable of fully expressing the UDFR data model.

The scope of this first version will include:

1. The specification and publicly-available documentation of:
   a. The export formats supported by this first version of UDFR
      i. These are formats the central registry can export for use by other services and tools
      ii. One of these formats will be the signature file used by DROID[5] (Digital Record Object Identification), a file format identification tool developed by the UK National Archives
   b. Any import formats supported by this first version of UDFR
      i. These are formats the central registry will accept that could be used to batch import records into the registry.

5 See http://sourceforge.net/projects/droid/

2. The design, development and testing of:
    a. A publicly accessible web-based user interface that can be used to search, browse, display and export registry records
    b. A web-based user interface that can be used to add and modify registry records
        i. The UDFR Governance Body Members will determine the extent and form of participation by the general public. Regardless, all UDFR content will be explicitly tagged to indicate the source of the information and the level of formal review.
    c. A publicly accessible web-based API that can be used by tools and services to query, retrieve and export registry records for local use
    d. The ability to export the signature file used by DROID
    e. A mechanism for ensuring unique UDFR identifiers
        i. If the UDFR identifiers are different from the PRONOM identifiers, the PRONOM identifiers will be also be maintained and available in the UDFR to provide ongoing support for the current PRONOM identifiers
    f. A mechanism for versioning record changes, including documenting the individual who made the changes

3. The import of the PRONOM records into the UDFR

The design of the system should account for the eventual inclusion of related documentation, specifications and test files in the registry. Future versions of the UDFR registry will build additional functionality onto this first version of the registry, as prioritized by the UDFR permanent government body.

IV. Management and oversight

Project oversight will be performed by CDL's Senior Manager for Digital Preservation Technology Project. The Senior Manager will communicate with the UDFR Governance Body and Technical Working Group for appropriate input. The Senior Manager will work closely with the CDL's Project Manager (Perry Willett) and Project Architect (to be hired) to ensure that the registry will meet the UDFR requirements. Project designs, plans, and deliverables will be shared with the UDFR Governance Body and/or the Technical Working Group, as appropriate.

V. Schedule and deliverables

The project duration is 12 months, starting January 1, 2010, and running through December 30, 2010. The project consists of 1 month of prototyping, 2 months of design and technical review, 8 months of development and 1 month of testing.

The major project milestones and deliverables are:

1. Month 1 [January 2010]. Staff recruitment, investigation and prototyping of a semantic wiki implementation to determine the feasibility of using a semantic wiki for the registry software.

2. Month 2 [February 2010]. In consultation with the UDFR Technical Working Group, make decision on the technology (semantic wiki, GDFR software, PRONOM software) to use as the technical base for the UDFR registry. While these activities will initially rely on electronic communication, a face-to-face meeting of the UDFR stakeholders will be organized at the end of month 2 to review and ratify design and implementation decisions.

3. Months 2-3 [February – March 2010]. Design of software enhancements for UDFR version 1.

4. Months 4-10 [April – October 2010]. Enhancement of software to meet the functional requirements for UDFR version 1.

5. Months 11-12 [October-November 2010]. Testing and documentation of software and APIs. The initial UDR operational production environment is assumed to be available in month 11 for testing purposes.

Ongoing user support and software maintenance and enhancement of the UDFR service is outside the scope of this proposal.

All UDFR software and documentation produced under the terms of this proposal will be made freely available under an open source license.

VI.Staffing

The UDFR technical development described in this proposal will be hosted at the California Digital Library (CDL) of the University of California. CDL will provide managerial oversight and administrative and technical support.

Project oversight will be directed by Stephen Abrams, Senior Manager for Digital Preservation Technology at CDL. Stephen Abrams is responsible for strategic planning, innovation, and operation of the CDL's digital curation infrastructure. He was the initiator of the Global Digital Format Registry (GDFR) project, leads the JHOVE2 project and participates in the UDFR Governance and Technical Working Groups.

Project Management and coordination will be provided by Perry Willett, CDL’s Digital Preservation Service Manager. Perry Willett will be responsible for high level coordination of the project with various stakeholders. Willett will also facilitate communication and reporting on project deliverables and milestones.

Patricia Cruse, UC Curation Center Director, CDL will provide **high-level coordination** and oversight of the project.

Two new staff will be hired for this project:

- **Project Architect.** This will be a senior person responsible for the detailed design and execution of the project. We will seek an individual with proven experience in the design and development of large web-based systems. Preference will be given to someone with proven experience using Web 2.0 technologies, including semantic wikis and web service APIs.

- **Project Developer.** This will be a mid-level position for implementing, testing, and documenting the registry system. We will seek an individual with proven experience in implementing sophisticated peer-to-peer information retrieval systems.

Appendix A UDFR Interim Governing Body Members

Every member of the UDFR Interim Governing Body is a member of the Governance Working Group. A subset are also members of the Technical Working Group. Pam Armstrong, Manager of the Library and Archives Canada's Digital Repository Services and Standards Office, is the chair of the Interim Body.

Governance Working Group

- British Library
- California Digital Library (CDL)
- Georgia Institute of Technology
- Harvard University Library
- Koninklijke Bibliotheeke
- Library and Archives Canada (LAC)
- Library of Congress
- National Archives and Records Administration (NARA)
- National Archives, UK (TNA)
- University of Illinois at Urbana-Champaign

Technical Working Group

- California Digital Library (CDL)
- Georgia Institute of Technology
- Harvard University Library
- Library and Archives Canada (LAC)
- National Archives, UK (TNA)
- University of Illinois at Urbana-Champaign

Appendix B Participants in GDFR, PRONOM and UDFR Initiatives

Andrew W. Mellon Foundation
Bibliothèque national de France
British Library
California Digital Library (CDL)
Cornell University
Corporation for National Research Initiatives
Digital Library Federation (DLF)
Drexel University
European Archive
Florida Center for Library Automation (FCLA)
Georgia Institute of Technology
German National Library
Harvard University
IBM Watson Research Center
Internet Architecture Board (IAB)
Internet Engineering Task Force (IETF)
Joint Information Systems Committee (JISC), UK
JSTOR
Koninklijke Bibliotheeke (KB)
Library and Archives Canada
Library of Congress
Los Alamos National Library
Massachusetts Institute of Technology (MIT)
National and University Library of Slovenia
National Archives and Records Administration (NARA)
National Archives, UK (TNA)
National Library of Australia
National Library of New Zealand
New York University
North Carolina State University
Online Computer Library Center (OCLC)
Oregon State University
Portico
Research Library Group (RLG)
Rutgers University
SAT Research Studio
Stanford University
Sweden Statens Arkiv
Tessella
TethersEnd Consulting
University of California Santa Barbara
University of Illinois at Urbana-Champaign
University of Maryland

University of Pennsylvania
University of Queensland
Upsala University Library
US Air Force Institute of Technology
US Government Printing Office
US General Services Administration
US National Aeronautics and Space Administration (NASA)
US National Guard
US National Institute of Technology and Standards (NIST)
WHOI/MBL