Unified Digital Format Registry (UDFR)

Final Report

2012-07-02

UC Curation Center
California Digital Library
University of California, Office of the President

Contents

1 Introduction

A deep understanding of digital formats is necessary to support the long-term preservation of digital assets, as it facilitates the preservation of the *information content* of those assets, rather than just their *bit stream representations*. A format is the set of syntactic and semantic rules that govern the mapping between information and the bits that represent that information. The Unified Digital Format Registry (UDFR), http://udfr.org/, is a new open source, semantically-enabled platform for the collection, long-term management, and dissemination of the significant properties of formats of interest to the preservation community[4]. The UDFR builds upon and "unifies" the function and holdings of two existing registry solutions: PRONOM, http://www.nationalarchives.gov.uk/PRONOM, from the UK National Archives since 2002; and GDFR (Global Digital Format Registry), http://gdfr.info/, from Harvard University since 2006. While these services rely on older relational and XML database technology, the UDFR uses a semantic database in which all information is represented in RDF form and exposed as Linked Data. Use of the UDFR is open to the public, although contribution or editing of information requires prior self-service account registration.

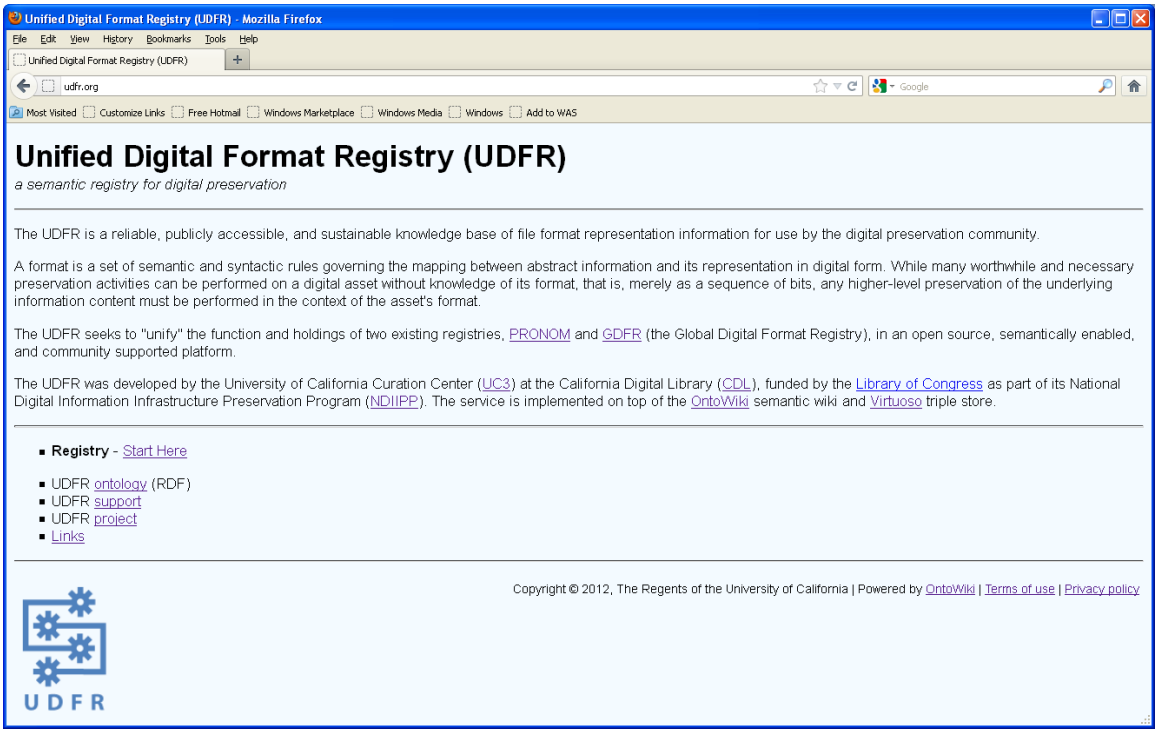
The UDFR was developed by the University of California Curation Center (UC3), http://www.cdlib.org/uc3, at the California Digital Library (CDL) with funding from the Library of Congress as part of its National Digital Information Infrastructure Preservation Program (NDIIPP), http://www.digitalpreservation.gov/.

The impetus for the UDFR project grew out of concerns raised by members of the international digital preservation community regarding the reliance on two previous format registry services: PRONOM and GDFR. While the UK National Archives asserts Open Government License (OGL) ownership over the *data* managed in the PRONOM registry, the intellectual property rights in the underlying technical *system* are retained in part by Tessella plc, which developed the system under contract with the Archives. Many preservation institutions and programs have raised concerns over this bifurcated ownership of PRONOM, especially with regard to long-term availability, and expressed a preference for a community supported, open source alternative.

The GDFR project intended to address this preference, but it was implemented using then state-of-the-art XML database technology. Since then, RDF-based semantic technologies have matured and many in the preservation community have expressed a

preference for a more forwarding-looking technological basis supporting a semantic database and the Linked Data accessibility. As a result, the GDFR never gained significant acceptance by the preservation community, holds a relatively modest set of format information, and must be considered a dormant initiative.

The UDFR is available at *http://udfr.org/*.



The registry itself is found at *http://udfr.org/registry/*. For automated access, a read-only SPARQL endpoint is available at http://udfr.org/ontowiki/sparql/. Additional links provide:

- UDFR ontology (RDF), *http://udfr.org/onto/onto.rdf*
- UDFR documentation, listserv, and presentations, *http://udfr.org/docs*
- UDFR project information, *http://udfr.org/project*
    - Proposal
    - Project team
    - Governance and technical working groups
    - Stakeholder meeting
    - Use cases and functional requirements
    - Risk assessment and mitigation strategies

- Technical architecture
- UDFR ontology
- Useful links, *http://udfr.org/links*

The *UDFR User's Guide* is available at http://udfr.org/docs/UDFR-Users-Guide-v1.0.0.pdf.

2 Proposal

Beginning in 2009, a series of ad hoc discussions within the preservation community crystallized around the idea of a new initiative to create a semantically-enabled, community supported, open source format registry that would support the union of function and holdings of PRONOM and GDFR. (At the instigation of Pam Armstrong, in her then current role as manager of digital repository services and standards at Libraries and Archives Canada, a number of these discussions occurred at meetings of the International Internet Preservation Consortium (IIPC). At no time, however, was the UDFR an official IIPC activity.) These discussions were organized under the purview of two informal working groups for governance and technology, including representatives from the following organizations.

Governance

- British Library
- California Digital Library
- Georgia Institute of Technology
- Harvard University
- Koninklijke Bibliotheek
- Library and Archives Canada
- Library of Congress
- National Archives [UK]
- National Archives and Records Administration
- University of Illinois, Urbana-Champaign

Technical

- California Digital Library
- Georgia Institute of Technology
- Harvard University
- Library and Archives Canada
- National Archives [UK]
- University of Illinois, Urbana-Champaign

These discussions led to a clear consensus as to the desirability of an independent, community supported open source format registry. The University of California

Curation Center (UC3), http://www.cdlib.org/uc3, one of five programmatic units at the California Digital Library (CDL), http://www.cdlib.org/, expressed an interest in providing the necessary development resources and in January 2010 provided a project proposal, http://udfr.org/project/UDFR-project-proposal.pdf, to the Library of Congress. The Library accepted the proposal as part of its National Digital Information Infrastructure and Preservation Program (NDIIPP), http://www.digitalpreservation.gov/.

The main deliverables from the proposal included:

- "*Specification and publicly-available documentation*"
- "*A publicly accessible web-based user interface that can be used to search, browse, display and export*"
- "*A publicly accessible web-based API that can be used by tools and services to query, retrieve and export*"
- "*Ability to export the signature file used by DROID*"
- "*A mechanism for ensuring unique UDFR identifiers*"
- "*Import of the PRONOM records*"

These requirements were later refined in consultation with the UDFR stakeholder community, as described in Section § 4.

Note that although initial discussions of the technical working group considered the idea of a distributed network of independent but cooperating registries, support for this feature was explicitly excluded from the final proposal. However, the design and implementation of the UDFR includes nothing that would preclude such support being added through follow-on development work.

The original project duration was one year. Due to difficulties in staffing, described in Section § 3, it was not possible to complete all project deliverables during that period. UC3 performed additional project work following the conclusion of funded performance.

3 Project team

UC3 project work commenced in January 2011 and completed in July 2012.

The UC3 project team was comprised of two full-time members, a project manager/ architect and developer. Existing UC3 staff filled other key roles as necessary.

- Lisa Dawn Colvin *project manager and architect* (full time)
- Abhishek Salve *developer* (full time)
- Stephen Abrams *project oversight and user advocate*
- Patricia Cruse *project oversight*
- John Kunze *noid consultant*
- Margaret Low *system administration*
- Mark Reyes *development and installation consultant*
- Marisa Strong *development consultant*

While the project manager/architect started at the beginning of the project in January 2011, UC3 experienced difficulty in filling the project developer role. An initial round of job postings and interviews failed to identify any acceptable candidates from the pool, primarily due to competition from local San Francisco-area commercial vendors offering more lucrative positions for developers with semantic web experience. UC3 then resorted to the use of a contract agency to supply the developer. The initial incumbent started in March 2011, one month later than intended in the original project plan, but was let go in May 2011 as the individual was not performing at an acceptable level. A replacement candidate, also provided through a contract agency, began work in July 2011. This transition midway through the project schedule caused significant delays in meeting the original development milestones.

The UC3 project was funded, and the project team supported, by the Library of Congress, as part of its National Digital Information Infrastructure Preservation Program (NDIIPP) initiative, *http://www.digitalpreservation.gov/*.

- Martha Anderson *NDIIPP director*
- Leslie Johnston *program officer*

The UDFR project received the full cooperation of the UK National Archives, http://www.nationalarchives.gov.uk/, and the Harvard University Library, *http://hul.harvard.edu/ois*.

- Andrea Goethals *Harvard University*
- Tim Gollins *National Archives*
- Tracey Powell *National Archives*
- Spencer Ross *National Archives*

As described in Section § 6, the UDFR is based on the OntoWiki semantic wiki platform. The developers of OntoWiki, the Agile Knowledge Engineering and Semantic Web (AKSW) research group at the University of Leipzig, *http://aksw.org/*, fully supported the efforts of the UC3 team.

- Philipp Frisch
- Norman Heino
- Sebastian Tramp

The support and contributions from the Library of Congress, UK National Archives, Harvard University, and AKSW were substantial and sustained, and greatly facilitated project work, as did the more informal participation of the UDFR stakeholder community.

4 Planning

In recognition of project's aggressive, one year scheduling, the UC3 project team early on engaged in a comprehensive risk analysis that led to the development of a number of mitigation strategies so that the team would be able to respond quickly should the need arise during the course of project work, *http://udfr.org/project/UDFR-risk-assessment.xlsx*. The identified risks fell into a number of high-level categories covering all significant aspects of project activities.

- *Staffing*. Hiring the full-time project manager/architect and developer with semantic expertise proved to be more difficult than anticipated, particularly the developer. This was exacerbated by the transition between developers three months into the project, as described in Section § 3. The final developer had excellent general development skills, but no specific semantic experience. In mitigation, the project team significantly modified the UDFR architecture to match the skill set. Previously, the intention was to use very low-level semantic tools and newly develop the majority of the application-level code. Since this option was no longer viable, the team selected the PHP-based OntoWiki semantic wiki platform in view of the developer's PHP experience and the comprehensive set of baseline application-level function provided by OntoWiki. The role of the project manager/architect, the only team member with significant semantic expertise, shifted to additionally encompass all ontology development and contributing to application testing. Similarly, responsibility for data transformation, i.e., cross-walking PRONOM XML to UDFR RDF, was

transferred to a UC3 manager in addition to his original oversight and user advocacy roles.

- *Schedule*. Recognizing the ambitious project goals for a one year schedule, the project team structured its work using agile development techniques. Development sprints were defined at a very granular level with the goal of a weekly stable build. At the end of each sprint, overall development priorities were reevaluated to select the scope for the next sprint. As schedule pressure grow in the last few months of the project, priority was given to activities performed by the two grant-funded project team members, the project manager/architect and developer. As a result, certain other activities, for example, user documentation, final production data loading, acceptance testing, and release, were postponed until after the end of funded project work. Additional UC3 staff were called upon to provide consultation and participate in building the operational environment and performing final acceptance testing and the production installation.

- *Technology*. The project team established an intention early on to rely on open source tools. Recognizing that the quality of such tools varies widely, preference was given to those tools with a proven track record and a vibrant developer and user community, leading to the selection of the OntoWiki/RDFauthor/Erfurt suite from the Agile Knowledge Engineering and Semantic Web (AKSW) research group at the University of Leipzig, *http://aksw.org/*, and the Virtuoso quadstore from OpenLink software, *http://virtuoso.openlinksw.com/rdf-quad-store/*. Although OntoWiki has proven to be an important and useful component of the UDFR architecture, it was not as robust and well-documented as original hoped for. In mitigation, the UDFR project team received excellent support from AKSW, who responded with alacrity to requests for information, correcting reported bugs, and adding requested functionality. In consequence, however, it was necessary to devote significantly more time than was originally anticipated to installation and unit testing, which compounded schedule slippage.

  While the original project plan set aside one month for evaluating various technological components for use by the UDFR, in actuality this process consumed closer to three months. Many of the open source tools under consideration were not fully mature and the process of installation and familiarization, let alone testing and evaluation, was time consuming. While the project team feels fully justified in the ultimate selection of the OntoWiki suite,

in hindsight it would have been preferable to have attempted to keep to the original schedule, perhaps by relying more on external reviews and consultation with expert users rather than extensive in-house evaluation.

While initial code development was performed on a Windows desktop machine, the preferred environment of the project developer, who did not have significant Unix experience, final service deployment was made to a Linux server to conform to standard UC3 hosting practice. Subtle differences in the functioning of the Virtuoso quadstore on these two platforms led to significant difficulties in diagnosing various error conditions that were not fully replicable across the development and production environments. Thus, the early decision to work in dual Windows/Linux environments in an effort to streamline development turned out to be somewhat counterproductive.

One unanticipated risk was that posed by the source code management of the OntoWiki/RDFauthor/Erfurt suite. The OntoWiki project moved between Google Code, *http://code.google.com/p/ontowiki/*, and Github, *https://github.com/AKSW/OntoWiki*, midway through the project timeline, necessitating changes to local deployment scripts. The original intention of the project team was to manage all local project code at the Bitbucket hosting site, *https://bitbucket.org/udfr/main*, for compatibility with other UC3 practices. However this required additional administrative overhead, so the UDFR code also was moved to Github, *https://github.com/UDFR*, for procedural simplicity. The OntoWiki/RDFauthor/Erfurt suite was subject to constant maintenance and development by AKSW during the entire period of UDFR project work. The project team found that they needed to schedule more frequent pulls from the source code repository than originally planned for in order to maintain a stable local development environment.

Another unanticipated risk was that posed by using a virtual machine as the server for the production UDFR service. The CDL had only very recently moved to the use of VMs and did not have significant experience in their provisioning and operation. A further complicating factor was that the CDL relies on different data centers and IT organizations for its development and production operations, and the procedures and level of support for VMs differed significantly. Also, the development VM was a server resource shared with other UC3 projects, complicating server configuration.

- *Community*. The initial level of community involvement was quite high as evidenced by the productive nature of the ad hoc working groups and the April 2011 invitational stakeholder meeting. In order to maintain this level of engagement the project team worked hard to ensure full transparency of all activities and decision making, relying on the UDFR-L mailing list, *UDFR-L@listserv.ucop.edu*, Bitbucket source code repository wiki and issue tracker, *http://bitbucket.org/udfr/main/wiki/Home*, and public presentations (see Section § 7). In practice, however, the level of subsequent community involvement declined markedly. In part this can be explained by the long gap in time between the early design decisions, circa April 2011, and the initial beta release for public evaluation in November 2011. While the original intention was to provide access to early prototypes, the difficulties experienced in development, described above, postponed the first release significantly. When the beta release became available for external testing in November 2011, significant feedback was received only from a small number of reviewers. In mitigation, the project team relied more heavily on the response of the user advocate team member.

In order to solidify fundamental design decisions, an invitational meeting of representative stakeholders was held at the Library of Congress in April 2011, *http://udfr.org/project/2011-04-13*. Meeting attendees included representatives from the following institutions:

- Bibliothèque national de France
- California Digital Library
- DataONE project/UCSB
- Deutsche Nationalbibliothek
- Ex Libris
- Family Search
- Florida Center for Library Automation
- Georgia Institute of Technology
- Government Printing Office [US]
- Harvard University
- Koninklijke Bibliotheek
- Library of Congress
- Los Alamos National Laboratory
- National Archives [UK]
- National Archives and Records Administration
- National Library of New Zealand
- New York University
- Open Planets Foundation / Nationaal Archief
- Tessella plc
- University of Pennsylvania
- Virginia Institute of Technology

The main topics on the meeting agenda were project background, uses cases and functional requirements, data modeling and ontologies, technical platform decisions, initial population, and community building. The major consensus decisions resulting from this meeting included:

- *Facts, not policies*. The scope of the information expressible by the UDFR data model is constrained to objective facts concerning formats. Policy statements and subjective evaluations, for example, an institutional preference for one format over another, is explicitly out of scope. (Should this decision be revisited in the future, the reliance of the UDFR data modeling on RDF ontologies would facilitate the process of adding additional representation information properties.)
- *Open contributor eligibility*. There are no prescriptive requirements for contributor eligibility other than providing minimum personal information: name, email address, and institutional affiliation and job title. Instead, the UDFR relies on strong provenance and complete change history at the level of each individual assertion.
- *Optional review*. All data managed in the UDFR should be subject to review on a per-assertion basis. The review status is made visible in the UDFR user interface. The recruitment of reviewers is outside the scope of project activities.
- *Opaque identifiers*. UDFR identifiers are opaque, rather than semantically meaningful.
- *No embargoes*. The implementation of the UDFR assumes that all of the format information managed in it is legally unencumbered. It does not need to support any form of embargoes on the public visibility of information.

In subsequent consultation with the stakeholder community, the UC3 project team made a number of significant policy decisions:

- All software resulting from project work is released under the terms of the GNU General Public License (GPL) for consistency with the licenses of third-party packages and software incorporated into the UDFR system.
- Data exported from PRONOM is made available under the terms of the UK Open Government License (OGL), *http://www.nationalarchives.gov.uk/doc/open-government-licence/*. The Internet Assigned Numbers Authority (IANA), the

source of MIME media type data, does not assert any particular terms of use for the various registries it makes available. Data submitted to the UDFR by contributors is made available under the terms of the Creative Commons Attribution (CC-BY) license, http://creativecommons.org/licenses/by/2.0/, whose terms must be agreed to by a contributor as a precondition for account registration.

5 Requirements

The public UDFR functional requirements, http://udfr.org/project/reqs.html, which were distilled from a set of use cases, http://udfr.org/project/twg/use-cases/, developed by the ad hoc working groups throughout 2009 and 2010, were the starting point for discussion at the two day, invitational stakeholders meeting at the Library of Congress in April 2011, convened in large part to ratify a set of final decisions on requirements and other aspects of the project planning. The overall intention of the final requirements supports the union of PRONOM and GDFR functionality. However, as the result of consensus decisions by the stakeholders group, several previously identified requirements were removed from consideration of the funded UDFR project.

- "*Support local copies of the registry*" with an "*ability to express value assessment[s] and priorities for using formats.*" The stakeholders agreed that the UDFR would not support a distributed model. The UDFR ontology does not support any means to distinguish between information that should and should not be shared. While there is nothing preventing an institution from installing a local copy of the UDFR and loading it with a full export, there are no explicit synchronization mechanisms supported by the UDFR codebase. Furthermore, the stakeholders agreed that the UDFR data modeling would focus on "facts, not policies."
- "*Support embargo metadata.*" The stakeholders agreed that the UDFR did not need to support data embargoes. All information submitted for inclusion in the UDFR is assumed to be unencumbered and is immediately visible to UDFR consumers.

Note that the while the scope of the original project included a requirement that "one of these [export] formats will be the signature file used by DROID (Digital Record Object Identification), a file format identification tool developed by the UK National Archives," this was genericized somewhat in the final requirements as "these [export] services should be usable by format identification tools external to the registry." While the

intention of the UDFR project team was to support the direct export of a DROID signature file, due to slippage of the overall project schedule this function was not implemented as part of the funded project work. Note, however, that the UDFR does manage all of the information expressible in the signature file, and it should be a straightforward process to translate the contents of an appropriate SPARQL query to conform to the signature file schema.

6 Architecture and technology

The UDFR architectural design evolved significantly through two phases. The first design phase was highly integrative in nature, composed of a granular set of digital library and semantic packages that were intended to be knitted together into a coherent application. (See the conceptual and concrete architectural diagrams at *http://udfr.org/project/arch.html*.) This initial design relied upon the following components:

- Drupal – content management system (CMS), *http://drupal.org/*
- Jena – semantic API, *http://jena.apache.org/*
- TDB – semantic triple store, *http://tw.rpi.edu/portal/TDB*
- Pubsubhubbub – syndication, *https://code.google.com/p/pubsubhubbub/*

After the original project developer was replaced, as described in Section § 3, this approach was abandoned in favor of a reliance on a more packaged solution, which was felt to be more expedient given that the remaining schedule period was not sufficient for the necessary integration work. Instead, the project team investigated the use of a packaged semantic wiki that would satisfy the project requirements. Note that the term "packaged" is relative. The final UDFR architecture is also composed of several components, but they explicitly support semantic wiki function, which would have had to have been developed under the first scheme. Explicit support for syndication, desired as a potential basis for the future development of a distributed network of independent but cooperating registries, was not included in the final architectural design.

Two main wiki packages were evaluated:

- Semantic Mediawiki (SMW), *http://semantic-mediawiki.org/*
- OntoWiki, *http://ontowiki.net/*

SMW is an extension of MediaWiki, http://www.mediawiki.org/, the platform for Wikipedia, http://www.wikipedia.org/. OntoWiki, on the other hand, is used by the EU-funded Linked Open Data (LOD2) project, http://lod2.eu/BlogPost/tag/linked-open-data, and is the basis for DBpedia, http://dbpedia.org/, the semantic re-expression of Wikipedia. In terms of wiki function, the two products are roughly equivalent. For semantic function, however, OntoWiki has several advantages. It supports RDF inferencing and partially supports OWL2. Additionally, OntoWiki uses a native RDF representation of the data managed in it, while SMW relies on a non-standard extension to its wiki markup language. Since any application such as the UDFR is inherently ephemeral, especially in distinction to its data, which is considered permanent, the UDFR project team wanted to facilitate long-term data continuity in the face of inevitable application obsolescence. Having the ability to export directly in a standard RDF serialization is therefore a benefit. While the SMW data could be transformed into equivalent RDF, this would require extra processing and care to ensure no transformational loss.

The final UDFR technology stack included the following components:

- OntoWiki – semantic wiki
- RDFauthor – RDFa form editor
- Erfurt – RDF API
- Virtuoso – SPARQL 1.1 quadstore
- Zend – PHP application MVC framework
- Apache httpd – webserver
- Noid – persistent identifier minter

OntoWiki, http://ontowiki.net/, supplies the main semantic wiki features, supporting the contribution, provenance, search, and display of semantic assertions. RDFauthor, https://github.com/AKSW/RDFauthor, is used for UI edit forms. Erfurt, http://aksw.org/Projects/Erfurt, is the API used as the interface between OntoWiki and the underlying SPARQL 1.1-enabled RDF quadstore, Virtuoso, http://www.openlinksw.com/dataspace/dav/ wiki/Main/VOSRDF. OntoWiki is a PHP application that runs as an extension of the Zend MVC framework, http://framework.zend.com/. The entire UDFR application runs in an Apache httpd webserver, http://projects.apache.org/projects/http_server.html. Persistent identifiers used to construct unique URIs for all RDF resources are minted using noid, https://wiki.ucop.edu/display/Curation/NOID. OntoWiki, RDFauthor, Erfurt, and Zend are all PHP packages, http://www.php.net/. Noid is a Perl module, http://www.perl.org/.

The UDFR project team performed UDFR-specific modifications in three main areas:

- Instance creation
- Support for technical review
- User profiles

The baseline OntoWiki adhered to the RDF principle of enabling "anyone to say anything about anything." Thus, at the time of instantiating a new RDF resource it was possible to add arbitrary RDF assertions. For the UDFR it is desirable to enforce tighter control over the instantiation process. Only those properties defined for a given class in the UDFR ontology can be added. This required substantial development effort.

Baseline OntoWiki does not support the concept of technical review. The UDFR project team had to modify OntoWiki to implement the review function, including the addition of a review checkbox on a per-assertion basis in the UI, and the inclusion of review status in the persistence layer.

While baseline OntoWiki does maintain user profiles, the granularity of the OntoWiki access control mechanism is such that either all or none of the profile information can be made visible. For UDFR it is desirable for a subset of profile information to be publicly visible. This subset includes the user name, title, and affiliation, which are revealed publicly as part of the provenance for each assertion. The UDFR collects and persists additional user information, most notably, an email address. While this is visible to UDFR administrators, it is properly suppressed from public view. This required the creation of a separate ontological model for the public user attributes.

The necessary modifications were spread across the OntoWiki, RDFauthor, and Erfurt system components. Consequently, the project team established three project repositories on GitHub forked from the main codelines.

- OntoWiki, *https://github.com/UDFR/OntoWiki*
  Forked from *https://github.com/AKSW/OntoWiki*

- RDFauthor, *https://github.com/UDFR/RDFauthor*
  Forked from *https://github.com/AKSW/RDFauthor*

- Erfurt, *https://github.com/UDFR/Erfurt*
  Forked from *https://github.com/AKSW/Erfurt*

The experience of working with OntoWiki, RDFauthor, and Erfurt was mixed, but ultimately successful. OntoWiki is still considered in a "pre-production release" state; that is, it is still at a pre "1.0" version. Many documented features did not originally work as described; other features that were fully or partially working were not properly documented. However, the UDFR project team received excellent technical support from the Agile Knowledge Engineering and Semantic Web (AKSW) research group at the University of Leipzig, http://aksw.org/. AKSW was extremely responsive to requests for information, bug fixes, and new features. In the end, all function necessary for the UDFR was fully implemented in the OntoWiki/RDFauthor/Erfurt code. The UDFR project team would like to acknowledge the technical support that it received from AKSW throughout the process.

7 Ontology

As defined in the functional requirements, the UDFR was intended to model the union of factual properties of the PRONOM and GDFR data models. (The most recent version of PRONOM supports policy information that is out of scope for UDFR.) At roughly the same time as UDFR project work was commencing, the UK National Archives announced an initiative to produce a Linked Data version of PRONOM. The UDFR was unable to rely on the PRONOM work as the basis for its own ontology, however, as the PRONOM development was occurring in parallel and the focus of that work covered only a subset of the full PRONOM data model. Nevertheless, the UDFR team did attempt to maintain consistency with the PRONOM to the fullest extent possible. However, the two teams made a number of different modeling choices. For example, in the UDFR enumerations are modeled as instances of an enumerate class; in PRONOM, enumerations are individual classes.

The PRONOM ontology of November 2011, *http://test.linkeddatapronom.nationalarchives.gov.uk/vocabulary/pronom-vocabulary.rdf*, defines 20 classes and 13 properties. The final UDFR ontology, *http://udfr.org/onto/onto.rdf*, defines 113 classes and 159 properties. To simply its definition, the UDFR ontology is organized with a hierarchical structure, with common groups of properties bundled together and associated with abstract classes for eventual inheritence. For example, the top of the hierarchy is the AbstractBase class, which is subclassed (rdfs:subclass) by concrete Agent, Digest, Holding, IPR, and Process classes and an AbstractProduct class, itself subclassed by concrete Document, File, Hardware, Media, and Software classes and AbstractFormat class, itself finally subclassed by

concrete CharacterEncoding, CompressionAlgorithm, and FileFormat classes. (See http://udfr.org/project/UDFR-class-hierarchy.pdf.)

Many UDFR properties have reasonable analogues in well-known ontologies such as Dublin Core (DC) and Friend of a Friend (foaf). However, in order to be able to enforce appropriate range and domain constraints properly, particularly during the instantiation of new resources, it was necessary for most of the properties to be explicitly defined in the UDFR ontology. To facilitate interoperation, however, various properties in the UDFR ontology are associated with other well-known ontologies through subtype relationships (rdfs:subProperty). For example, the UDFR "documentAuthor" is defined as a sub-property of Dublin Core "creator."

8 Data loads

The UDFR was seeded with data from two sources:

- MIME media types
- PRONOM

The MIME media types were exported from an aggregated registry on Appspot, *http://mediatypes.appspot.com/*, which is "routinely scrapped from IANA [*http://www.iana.org/assignments/media-types*] using code in the mediatypes Google Code project [subsequently moved to GitHub, *https://github.com/edsu/mediatypes*]," on February 22, 2012. This data includes:

Count	***MIME type***
809	application/*
125	audio/*
39	image/
19	message/*
14	model/*
14	multipart/*
51	text/*
56	video/*
1,127	

Format data from the PRONOM registry, *http://www.nationalarchives.gov.uk/PRONOM*,

was exported on February 22, 2012, using a Python script provided by the National Archives, http://www.nationalarchives.gov.uk/. This data includes:

Count	*Ontological class*
846	file formats
28	character encodings
17	compression algorithms
1,237	Identifiers
548	external signatures
494	internal signatures
71	MIME types (not in IANA/Appspot)
156	agents
268	software packages
2,080	software processes
23	intellectual property rights (IPR) statements
217	relationships
5,985	

The UDFR project team wishes to acknowledge the generous technical support of the National Archives throughout this process.

The Appspot MIME data export was in RDF using the XML serialization. The PRONOM export was in XML using a PRONOM-specific schema. In both instances the UDFR project team applied Perl scripts to perform the necessary crosswalk to the UDFR ontology expressed as RDF in the N-triples serialization. The data was then bulk uploaded using a curl script. In all, these two imports correspond to 42,617 RDF triples, with 7,341 unique subjects and 4,487 unique objects.

9 Community engagement

The UC3 project team initially selected Bitbucket, http://bitbucket.org/, as the public code hosting site, and established a wiki for community communications at http://bitbucket.org/udfr/main/wiki/Home. It later became useful to move the UDFR code and ontologies to GitHub, http://github.com/UDFR, in order to streamline integration of key technological components of the UDFR application. For convenient retrieval, all information from the Bitbucket wiki was migrated to the UDFR web site,

http://udfr.org/project, and the use of Bitbucket was deprecated.

A mailing list, UDFR-L@listserv.ucop.edu, was established for direct communication with the UDFR user community. An online archive is available at http://listserv.ucop.edu/cgi-bin/wa.exe?A0=UDFR-L.

The UDFR project team made a number of public presentations during the course of the project covering technical, operational, and governance issues.

- "UDFR: A semantic registry for format representation information," *DLF Forum*, Baltimore, October 31—November 2, 2011, http://udfr.org/docs/DLF-2011-UDFR-A-Semantic-Registry-for-Format-Representation-Information-v1.pptx
- "Unified Digital Format Registry (UDFR): Overview and steps to an operational registry," *PASIG 2012*, Austin, January 11-13, 2012, http://udfr.org/docs/PASIG-2012-UDFR-v03.pptx
- "Unified Digital Format Registry (UDFR): Understanding the system and service," *IIPC General Assembly*, Library of Congress, April 31—May 4, 2012, http://udfr.org/docs/IIPC-2012-UDFR-community-meeting-v07.pptx
- Unified Digital Format Registry workshop, *Digital Preservation 2012*, Washington, July 24-26, 2012

The workshop at the IIPC General Assembly at the Library of Congress in May 2012 was the most comprehensive public presentation on the UDFR. The five hour session covered all major areas of the UDFR project and service: background, demonstration of features, technology and architecture, code walkthrough, ontological modeling, administrative procedures, and community building and next steps. For more information, see http://udfr.org/docs/UDFR-community-meeting-2012-05-04.pdf and http://udfr.org/docs/IIPC-2012-UDFR-community-meeting-v07.pptx.

Discussion during the workshop included a suggestion to modify the UDFR data model so that the external signatures (e.g., file extensions) imported from PRONOM would be de-duplicated. This proposal was posted on the UDFR-L mailing list for wider public comment, http://listserv.ucop.edu/cgi-bin/wa.exe?A2=ind1205&L=UDFR-L&P=55, where the proposal received favorable response. The change was implemented, with the result that the 1,006 individual file extension instances in PRONOM were coalesced into 548 unique external signatures. During the subsequent reloading of modified PRONOM

data, an error in the original RDF was also corrected. The identifier namespace type resources were defined using an incorrectly form URL, *udfr:identifierNamespace* instead of *udfrs:identifierNamespaceType*. This error, reported during the beta evaluation period, had affected all 1,237 identifiers.

The workshop at the Digital Preservation 2012, the annual NDIIPP/NDSA meeting, focused on discussions of long-term governance, administration, and operation.

The initial UDFR stakeholder community was drawn from institutions actively involved with the GDFR, PRONOM, and UDFR initiatives.

- Air Force Institute of Technology [US]
- Andrew W. Mellon Foundation
- Bibliothèque national de France
- British Library
- California Digital Library
- Cornell University
- Corporation for National Research Initiatives
- Digital Library Federation
- Drexel University
- European Archive
- Florida Center for Library Automation
- General Services Administration [US]
- Georgia Institute of Technology
- German National Library
- Government Printing Office [US]
- Harvard University
- IBM Watson Research Center
- Internet Architecture Board
- Internet Engineering Task Force
- Joint Information Systems Committee
- National and University Library of Slovenia
- National Archives [UK]
- National Archives and Records Administration
- National Guard [US]
- National Institute of Technology and Standards [US]
- National Library of Australia
- National Library of New Zealand
- New York University
- North Carolina State University
- OCLC
- Oregon State University
- Portico
- Rutgers University
- SAT Research Studio
- Stanford University
- Statens Arkiv
- Tessella
- TethersEnd Consulting
- University of California, Santa Barbara
- University of Illinois, Urbana-Champaign
- University of Maryland

- JSTOR
- Koninklijke Bibliotheek
- Library and Archives Canada
- Library of Congress
- Los Alamos National Library
- Massachusetts Institute of Technology
- NASA
- University of Pennsylvania
- University of Queensland
- Uppsala University Library
- Woods Hole Oceanographic Institution/Marine Biological Laboratory

10 Conclusion

The UDFR was released for production use in July 2012, http://udfr.org/, meeting all major functional requirements, except for the ability to export a PRONOM signature file. (As previously discussed, all of the necessary information for the file is found in the UDFR, but the specific form of the file cannot be generated directly.) The UDFR supports the combined function found in PRONOM and the GDFR and initial holds a full February 2012 export of the IANA MIME type and PRONOM registries.

The implementation in terms of a semantic wiki is a major advance over the older relational and XML database technology used by PRONOM and the GDFR. It positions the UDFR to interoperate smoothly with the evolving semantic web and the growing ecosystem of Linked Data-aware applications. The use of a semantic platform did add to the complexity of the development process, however, due to the open source tools not being as robust as might be desired and the difficulty in finding appropriately-experienced staff. These factors contributed to schedule pressure, with the result that several important project activities were not performed until after the end of funded work.

Through engagement with the UDFR stakeholder and user communities, several significant follow-on activities have been identified.

- *Replication*. Support for dynamic replication between multiple instances was considered out of scope for the funded project work. However, nothing in the design or implementation of the UDFR would preclude the extension for replication in the future.
- *Additional data sources*. In addition to encouraging the individual contribution of format information to the UDFR, it would be beneficial to investigate sources

that could be bulk imported. Two candidate sources are the Library of Congress's Sustainability of Digital Formats website, http://www.digitalpreservation.gov/formats/, and the IT History Society's hardware database, http://www.ithistory.org/hardware/hardware-name.php.

- *Reviewers*. While the UDFR supports technical review of individual assertions, the recruitment and training of suitable reviewers was out of scope for funded project work. The long-term success of the UDFR is highly dependent on its being perceived as a reliable source of important format representation information. Having that information subject to some level of review by recognized experts is therefore an important component of service trustworthiness. One potential model is provided by the practices of the Internet Engineering Task Force (IETF) with respect to Requests for Comments (RFC), in which proposals are accepted from the public (analogous to public contribution to the UDFR), subject to public review (analogous to the UDFR comment facility), and then layers of increasingly anonymous review by IETF technical exports (analogous to UDFR technical review).
- *Permanent operational home*. Under the terms of the original proposal to the Library of Congress, the California Digital Library (CDL) will continue to operate the UDFR for one year until a more permanent operational home can be found. CDL may offer to continue on in that role, but only if accompanied by some form of partial cost recovery for operational support.
- *Permanent governance*. The UDFR was always conceived of as a community supported resource. Control over its long-term direction with regard to policy, maintenance, operation, and enhancement must then be provided by the community. It is therefore important to define some lightweight process under which this governance can be exercised.

References

[1] Apache Software Foundation, *Apache HTTP Server* <http://projects.apache.org/projects/http_server.html>.

[2] California Digital Library, *California Digital Library (CDL)* <http://www.cdlib.org/>

[3] California Digital Library, *Noid: Nice Opaque Identifier (Minter and Name Resolver)* <https://wiki.ucop.edu/display/Curation/NOID>.

[4] California Digital Library, *Unified Digital Format Registry (UDFR)* <http://udfr.org/>.

[5] California Digital Library, *University of California Curation Center* (*UC3)* <http://www.cdlib.org/uc3>.

[6] Harvard University Library, *Global Digital Format Registry (GDFR)* <http://gdfr.info/>.

[7] Internet Assigned Numbers Authority, *MIME Media Types* <http://www.iana.org/assignments/media-types/index.html>.

[8] Library of Congress, *Digital Preservation* <http://www.digitalpreservation.gov/>.

[9] *Linked Data – Connect Distributed Data across the Web* <http://linkeddata.org/>.

[10] *Media-types* <http://mediatypes.appspot.com/>.

[11] National Archives [UK], *File profiling tool (DROID)* <http://www.nationalarchives.gov.uk/information-management/our-services/dc-file-profiling-tool.htm >.

[12] National Archives [UK], *PRONOM* <http://nationalarchives.gov.uk/PRONOM/Default.aspx>.

[13] National Archives [UK], *PRONOM Vocabulary Specification: DRAFT*, October 26, 2011 <http://test.linkeddatapronom.nationalarchives.gov.uk/vocabulary/pronom-vocabulary.htm>.

[14] Open Link Software, *Virtuoso RDF* <http://www.openlinksw.com/dataspace/dav/wiki/Main/VOSRDF>

[15] PHP Group, *PHP – Hypertext Preprocessor* <http://www.php.net/>.

[16] Prud'hommeaux, Eric, and Andy Seaborne, eds., *SPARQL Query Language for*

RDF, W3C Recommendation, January 15, 2008 <http://www.w3.org/TR/rdf-sparql-query/>.

[17] *Semantic MediaWiki* <http://semantic-mediawiki.org/>.

[18] University of Leipzig, *Agile Knowledge Engineering and Semantic Web (AKSW)* <http://aksw.org/About>.

[19] University of Leipzig, *Erfurt* <http://aksw.org/Projects/Erfurt>.

[20] University of Leipzig, *OntoWiki* <http://ontowiki.net/>.

[21] University of Leipzig, *RDFauthor* <https://github.com/AKSW/RDFauthor>.

[22] Wikipedia, *Linked data* <http://en.wikipedia.org/wiki/Linked_data>.

[23] World Wide Web Consortium, *N-Triples*, W3C RDF Core WG internal Working Draft <http://www.w3.org/2001/sw/RDFCore/ntriples/>. See also <http://www.w3.org/TR/rdf-testcases/#ntriples>.

[24] World Wide Web Consortium, *OWL 2 Web Ontology Language: Document Overview* <http://www.w3.org/TR/owl2-overview/>.

[25] World Wide Web Consortium, *Resource Description Framework (RDF)* <http://www.w3.org/RDF/>.

[26] Zend Technologies, Ltd., *Zend Framework* <http://framework.zend.com/>.