Unified Digital Format Registry (UDFR) Stakeholder Meeting

Library of Congress, April 13-14, 2011

The UDFR project seeks to develop a community-supported, open source, on-line registry of file format representation information of interest for purposes of digital preservation, curation, and repository operation. In order to ensure that the UDFR meets the needs of its diverse user constituencies, an invitational stakeholder meeting will be held on April 13-14 at the Library of Congress in Washington, DC, US.

Outcomes

The desired outcomes of this stakeholder meeting include:

- Appropriate scoping of the UDFR project through agreement on functional and non-functional requirements.
- Agreement on the data modeling process and the resulting ontology.
- Agreement on key technology decisions.
- Agreement on project plan and schedule.
- Groundwork for fostering of an active stakeholder community providing for administrative and technical continuity of the UDFR as an ongoing service following the conclusion of the funded project.

Agenda

Wednesday, April 12, 2011

Time	Topic
09:00 – 09:20	Welcome and introductions
09:20 – 09:30	Review of objectives and agenda
09:30 – 10:00	Project background
10:00 – 10:30	Use cases and functional requirements
10:30 – 11:00	*Break*
11:00 – 11:30	Use cases and functional requirements (*continued*)
11:30 – 12:30	Data modeling and ontology
12:30 – 13:30	*Lunch*
13:30 – 14:30	Data modeling and ontology (*continued*)
14:30 – 15:00	Technical architecture
15:00 – 15:30	Break
15:30 – 16:30	Technical platform decisions
16:30 – 17:00	Questions and discussion
17:00	*Adjourn*

Thursday, April 13, 2011

Time	Topic
09:00 – 09:30	Project schedule
09:30 – 10:15	Initial population of UDFR
10:15 – 10:45	Community building
10:45 – 11:15	*Break*
11:15 – 12:30	Community building (*continued*)
12:30 – 13:00	Follow-up planning
13:00	*Adjourn*

Background

Information on the UDFR project is available at <http://www.udfr.org/>. The function and design of the UDFR is heavily reliant on the prior work of the PRONOM registry <http://www.nationalarchives.gov.uk/PRONOM/Default.aspx> and the GDFR project <http://www.gdfr.info/>.

Additional document relating to the design and implementation of the UDFR are available on the project wiki at <http://udfr.org/wiki/index.php/UDFR_Stakeholder_Meeting_Documents>.

Requirements

We assume that the use cases and requirements developed by the stakeholder community over the past years are fairly complete and reasonably non-controversial. However, there are a small number of important questions around which there is not yet clear consensus.

- Scope of data modeling: what subsets of PRONOM and GDFR modeling are necessary and desirable?
- What are the criteria for contributor eligibility? Anonymous? Public, but known? Self-nominated, but vetted? Invited?
- What level of technical review should/will contributed information be subjected, and by whom?
- Should the UDFR format identifiers be opaque or transparent?
- Does the GDFR distinction between format families and formats that are members of those families make sense?
  (See <http://gdfr.info/docs/GDFR-Format-Model-and-Relationship-1_0_7.pdf>)
- Should the registry support actionable inheritance of properties? For example, should BWF automatically inherit all properties defined for "generic" WAVE?
- Is there a useful distinction between "facts" and "institutional policy"?
  (See, for example, Open Planets Foundation, *A New Registry for Digital Preservation: Conceptual Overview*, January 2011 <http://www.openplanetsfoundation.org/new-registry-digital-

preservation-conceptual-overview>)
- What is the proper granularity for provenance and technical review, per-property or per-aggregate entity (e.g., format, agent, document, etc.)
- Does there need to be an explicit grant of license by content contributors (e.g. CC0)?
- Replication model: master/slave(s) or peer-to-peer?
- What is the ongoing relationship between PRONOM and UDFR?

Data Modeling and Ontology

UDFR data modeling and ontology development will be discussed in terms of:

- Goals and guiding principles
- High-level topics
    - Terms
    - Identifiers
    - Conceptual entities
    - Controlled vocabularies
    - Use cases: format lifecycle, provenance, governance
    - Modularization

Architecture

Working assumptions regarding the UDFR's technical architecture:

- Web <http://www.w3.org/TR/webarch/>.
- REST <http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm >.
- Linked data <http://linkeddata.org/>.
- MVC <http://en.wikipedia.org/wiki/Model%E2%80%93view%E2%80%93controller >.

The UDFR project team anticipates the bulk of the project will consist of integrating together existing open source components supporting necessary function, particularly with regard to:

- Triple/quad/quint store.
- Semantic wiki.
- Management of physical assets.
- Index/search.
- Replication.