Building a robust knowledge base for digital formats

John Mark Ockerbloom

DLF Spring Forum

April 21, 2004

Why this talk?

- Figure out how to build such a knowledge base
  - Raise a lot of questions about how to do it
  - I'll propose some answers in this talk, but I'd like you to be thinking about the questions
- Show a prototype designed to help answer some of these questions
  - Fred (Format REgistry Demonstration)
- Invite you to participate in the process

- (And thank sponsors of this effort, in particular the DLF and the Mellon Foundation)

Why a format registry?

- Detailed format knowledge needed for many digital repository and preservation activities
- But the knowledge is now being assembled
  - » Piecemeal (no one place to look, or even start looking)
  - » Incompletely (much format information not public)
  - » Redundantly (lots of projects need the same info)
- Some registries already exist…
  - » IANA's MIMEregistry is the most widely used
  - » Some organizations (PRO, Dspace, Apple) have registries for their own constituency or software
- …but have insufficient breadth, depth, or participation for the needs of preservation

Some challenges for a global format registry

- What information does it really need to manage?
- How is this information going to be accumulated, mantained, used?
- How can it get both high quantity and good quality of information?
- How will it interact with other people, systems?
- Practical issues: How will it be governed, supported?

Why a format registry prototype?

- "The best is the enemy of the good"
  - Get rid of the paralysis induced by having to get it right the first time
  - For now, forget the idea that a single registry has to do everything
- Provide a more concrete basis for discussion and planning of global digital format registry system
- Test ideas about data model, service model
  - Starting with a data model prepared by Stephen Abrams based on previous GDFR meetings
  - Starting with a simple interactive Web site for viewing and entering information. (Other interfaces to follow.)
- Get some experience collecting, managing format description data
- Plan for a persistent, sustained, more authoritative registry system

What should a registry do?

- **Name formats**
  - » **Give them unambiguous, ubiquitous names**
    - **Analogy: naming authority databases**
  - » **and/or provide a mechanism to assign such names**
    - **Analogy: DNS**
  - » **Ideally, the naming should have global reach, comprehensiveness**
- **Document formats, their identification, and their use:**
  - » **Describe: Document format characteristics, specs, risks**
  - » **Identify: Assist in the determination, validation of object formats**
  - » **Use: Point to tools and methods to render, characterize,extract information from, and convert objects in that format**
- **Registry can contain original information, or pointers**

How are formats described?

- **Syntactically**: What the bytes look like
  - E.g. BSDL, grammars
  - Verification: Signatures and other “magic number” checking
- **Structurally**: What data structures are present
  - E.g. DIDL, ASN.1 declarations, many specification documents
  - May assume particular underlying syntax (e..g XML Schemas), or abstract away from it (e.g. Dublin Core and OAIS definitions)
- **Functionally**: What can be done with the format
  - E.g. TOM: Object-oriented approach (attributes, methods…)
  - E.g. PRONOM: Programs that use, maintain, produce the format
- **Practically**: Effective use and sustainability of format
  - Support, quality, IP restrictions, etc.
  - (Many issues covered in LC talk from Fall 2003 Forum)
- We care about all of these
  - But might not be able to address them all perfectly
  - They also matter for abstract “formats”: Scope creep?

Where does the information come from?

- From the designers of the formats
  - » You need to find them
  - » They need the time to provide the documentation
  - » They need to trust you
  - » They might not be the most objective about utility
- From the researches of the registry maintainers
  - » They need the time and expertise to seek it out and record it
- From third parties (developers, researchers, the public)
  - » They need to be able and willing to contribute
  - » There needs to be some way to evaluate the validity and quality of the information provided
    - Either to admit it, or to rate it
- A comprehensive system needs all of these sources

What can we learn from MIME?

- **Simple naming conventions, registration procedure can be widely used**
  - **Sponsorship by IANA also helps**
- **Less information than we'd like**
  - **Many more specialized formats not covered**
  - **Many listed formats not covered adequately**
    - » **Maybe because of limitations on who can register**
    - » **Specialized format variants not well covered; maybe format relations need to be better supported**
- **Naming needed some work**
  - **"Experimental" tree got messy**
  - **"Vendor" tree helped enlarge repertoire, still incomplete**

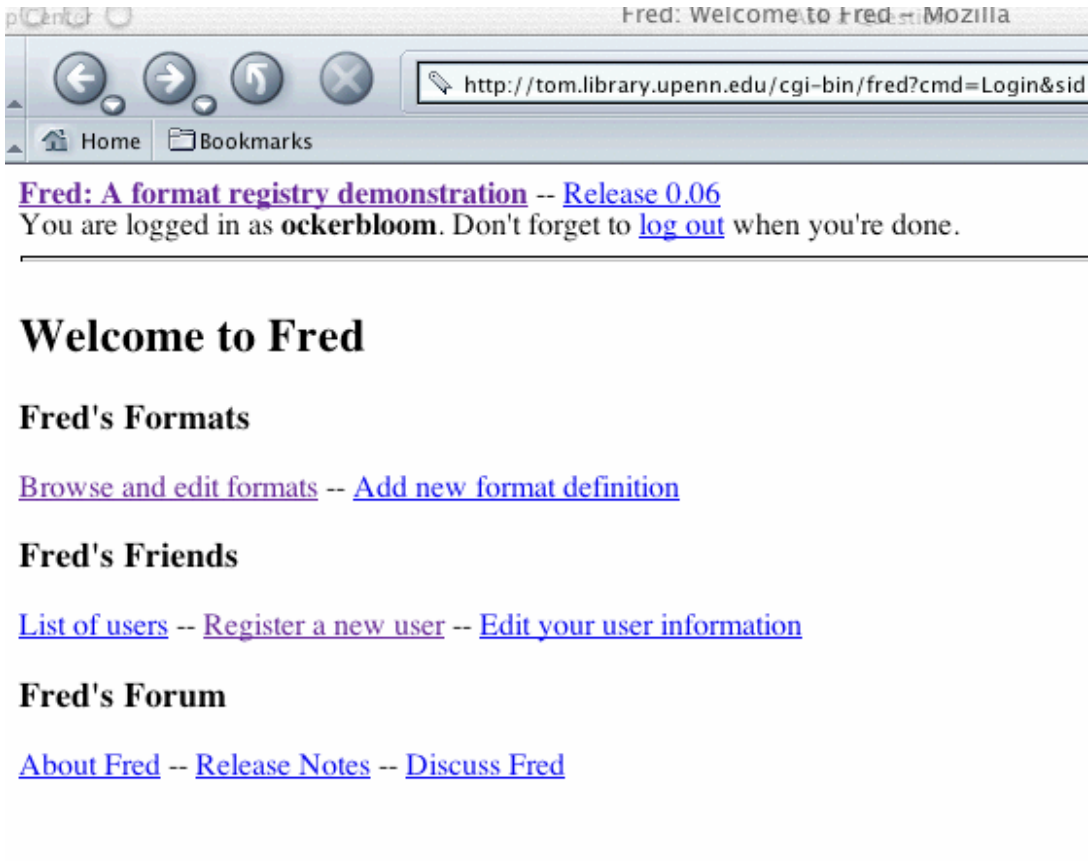
What can we learn from Wikipedia?

- **Over time, you can pick up a lot from a large, interested user community**
  - **And even slowly progress towards more reliable information**
  - **And develop useful structures, conventions over time**
- **Trustworthiness still a concern**
  - **Attribution, histories, discussion, all help evaluate and maintain trustworthiness**
  - **Can also interact with more formal peer review**
    - » **Not used much in Wikipedia (in part because there are other peer-reviewed encyclopedias)**
    - » **Used elsewhere for monographs, open source software**
- **Interlinking (within, into, and out of) a key strength**
- **Note: A format registry system would have a smaller user base**
  - **(though ßprobably more expert and motivated on average)**

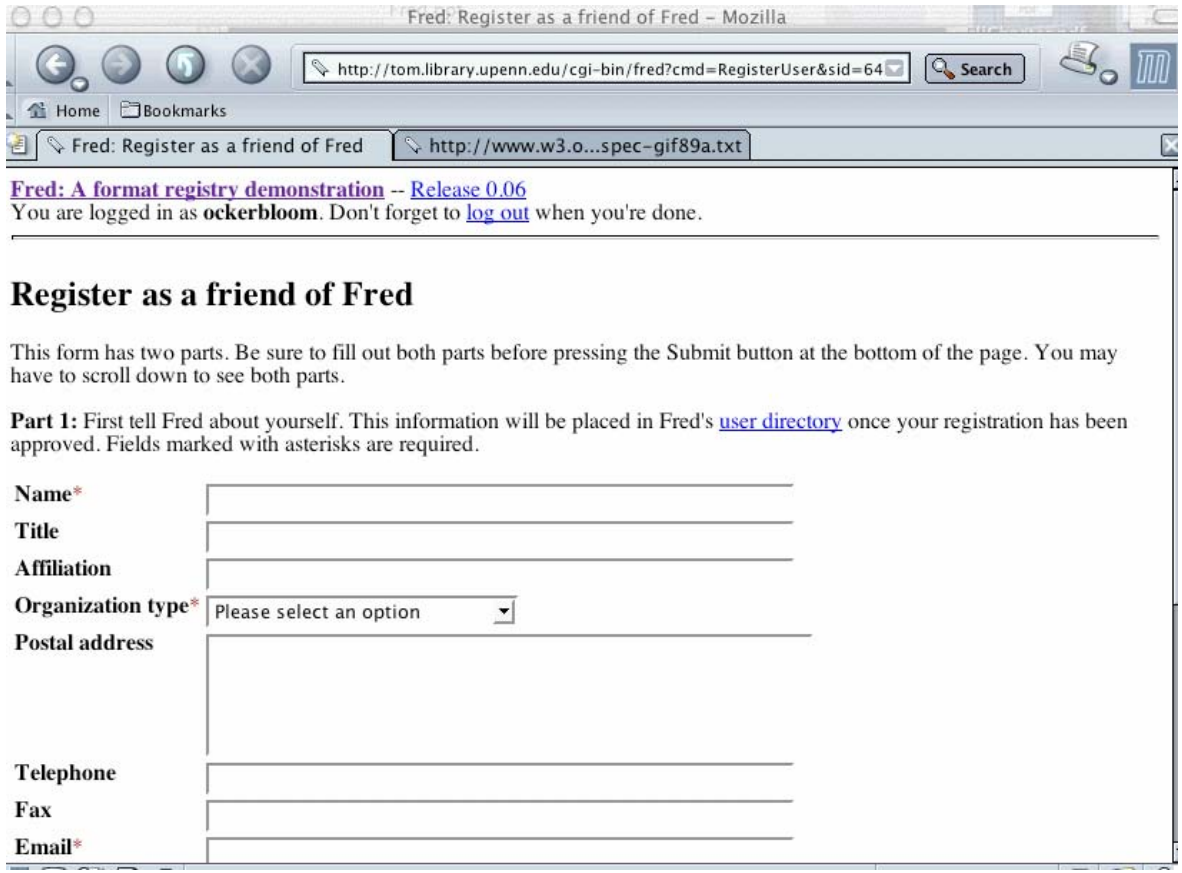
How have we designed Fred to work as a testbed?

- Make it easy to evolve
  - » Start with a basic data model based on GDFR discussions
  - » Represent in XML, render using generic, data-driven display and editing modules (also to be used with TOM)
  - » Makes for easy export, import, remapping
- Make it accommodate a continuum of control, structure
  - » Core in structured form with strict access control
    - But: format authors can open up permissions if desired
  - » Each format also has unstructured Wiki discussion
- Promiscuously link with related information resources
  - » Aliases link to other “registry” systems (e.g. TOM, MIME)
  - » Within registry system, format relations and repository-specific namespaces may encourage network growth
- Make it highly reflective and participatory
  - » Linked Wikis let users comment on, improve, information
  - » Full editing , discussion, review history maintained

Quick tour


Fred: Welcome to Fred -- Mozilla
http://tom.library.upenn.edu/cgi-bin/fred?cmd=Login&sid
Home
Bookmarks
Fred: A format registry demonstration -- Release 0.06
You are logged in as ockerbloom. Don't forget to log out when you're done.
Welcome to Fred
Fred's Formats
Browse and edit formats -- Add new format definition
Fred's Friends
List of users -- Register a new user -- Edit your user information
Fred's Forum
About Fred -- Release Notes -- Discuss Fred

Quick tour


Fred: Register as a friend of Fred - Mozilla
http://tom.library.upenn.edu/cgi-bin/fred?cmd=RegisterUser&sid=64
Search
Home
Bookmarks
Fred: Register as a friend of Fred
http://www.w3.o...spec-gif89a.txt
Fred: A format registry demonstration -- Release 0.06
You are logged in as ockerbloom. Don't forget to log out when you're done.
Register as a friend of Fred
This form has two parts. Be sure to fill out both parts before pressing the Submit button at the bottom of the page. You may have to scroll down to see both parts.
Part 1: First tell Fred about yourself. This information will be placed in Fred's user directory once your registration has been approved. Fields marked with asterisks are required.
Name*
Title
Affiliation
Organization type*
Please select an option
Postal address
Telephone
Fax
Email*

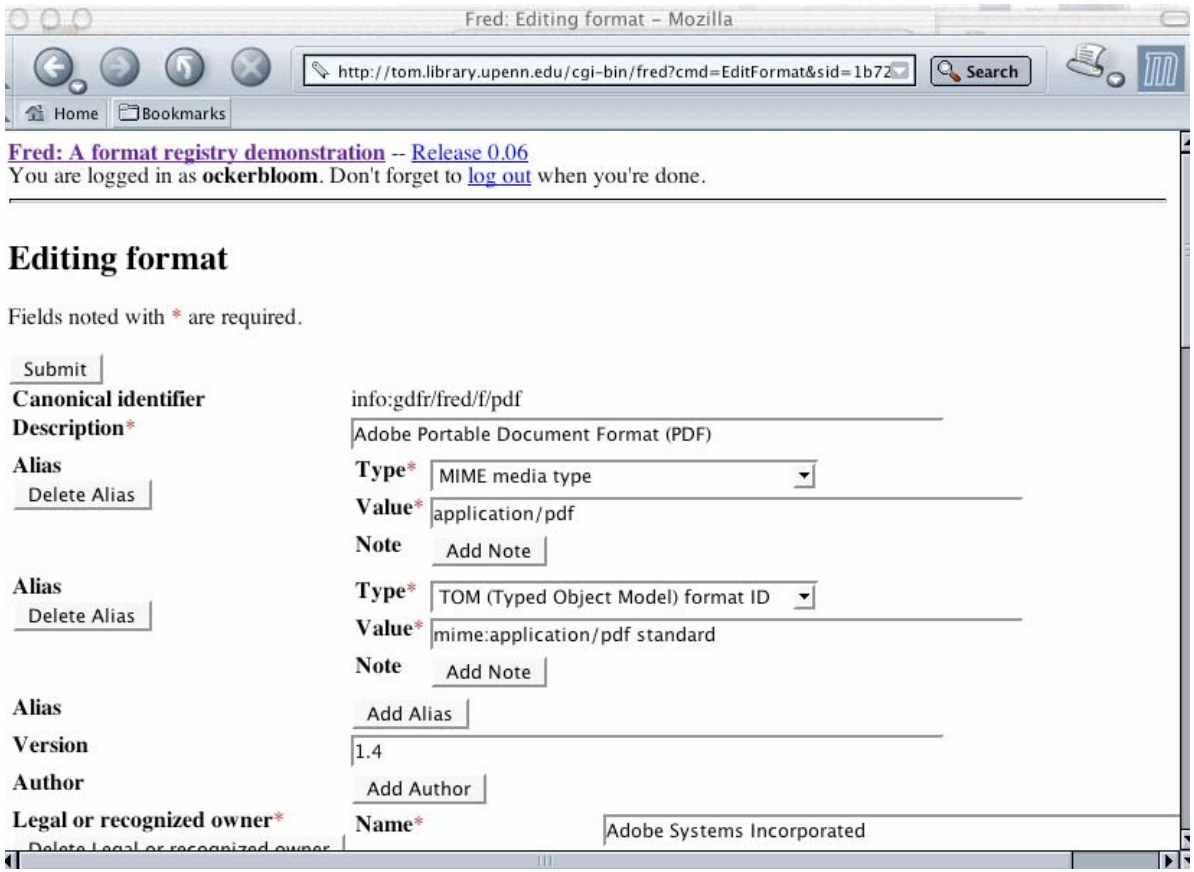
Quick tour


Fred: Format: info:gdfr/fred/f/pdf - Mozilla
http://tom.library.upenn.edu/cgi-bin/fred?cmd=ShowFormat&sid=1b7
Search
Home
Bookmarks
Fred: A format registry demonstration -- Release 0.06
You are logged in as ockerbloom. Don't forget to log out when you're done.
Format: info:gdfr/fred/f/pdf
See permissions -- See edit history -- Discuss -- Edit
Canonical identifier info:gdfr/fred/f/pdf
Description Adobe Portable Document Format (PDF)
Alias MIME type application/pdf
Alias TOM format mime:application/pdf standard
Version 1.4
Legal or recognized owner
Name Adobe Systems Incorporated
Organization type Commercial (for-profit) entity
Web site http://www.adobe.com/
Description last modified 6:38:25 PM, on March 23, 2004 (GMT)
Specification
Document title PDF Reference: Adobe Portable Document Format, Version 1.4
Document type Manual
Edition 3rd edition
Access regime Unrestricted access
Identifier http://partners.adobe.com/asn/acrobat/docs/File_Format_Specifications/PDFReference.pdf
Identifier
Type ISBN
Value 0-201-75839-3
Last modified 9:00:02 PM, on April 16, 2004 (GMT)

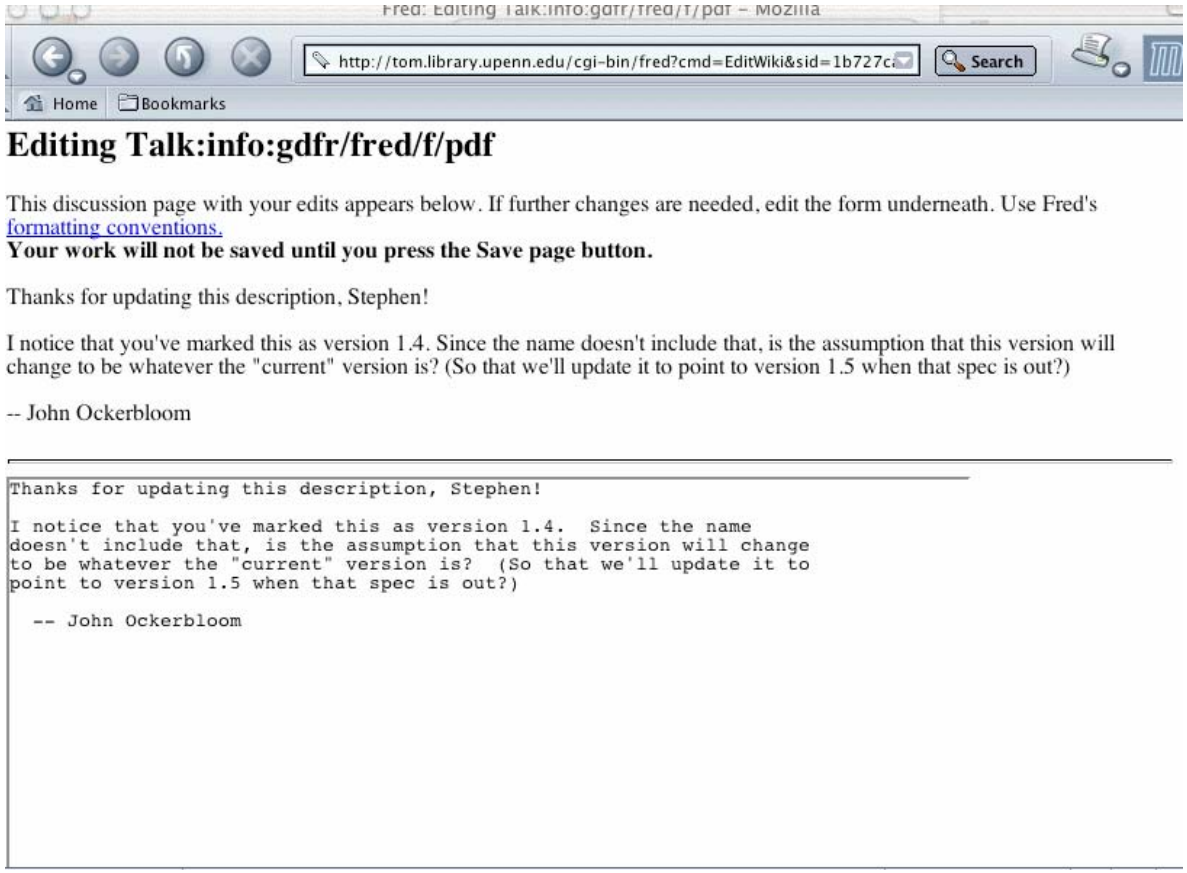
Quick tour


Fred: Format: info:gdfr/fred/f/pdf - Mozilla
http://tom.library.upenn.edu/cgi-bin/fred?cmd=ShowFormatHistory&s
Home
Bookmarks
Fred: A format registry demonstration -- Release 0.06
You are logged in as ockerbloom. Don't forget to log out when you're done.
Format: info:gdfr/fred/f/pdf
Editing history
Select a line to see details of the revision. Lines in bold mark substantial technical changes.
21:00, 16 Apr 2004, by stephen: [Full review]
19:40, 23 Mar 2004, by [see details]: Added a pointer to the TOM format.
18:38, 23 Mar 2004, by [see details]: Bare-bones entry for starters

Quick tour


Fred: Editing format - Mozilla
http://tom.library.upenn.edu/cgi-bin/fred?cmd=EditFormat&sid=1b72
Search
Home
Bookmarks
Fred: A format registry demonstration -- Release 0.06
You are logged in as ockerbloom. Don't forget to log out when you're done.
Editing format
Fields noted with * are required.
Submit
Canonical identifier
info:gdfr/fred/f/pdf
Description*
Adobe Portable Document Format (PDF)
Alias
Delete Alias
Type*
MIME media type
Value*
application/pdf
Note
Add Note
Alias
Delete Alias
Type*
TOM (Typed Object Model) format ID
Value*
mime:application/pdf standard
Note
Add Note
Alias
Add Alias
Version
1.4
Author
Add Author
Legal or recognized owner*
Name*
Adobe Systems Incorporated

Quick tour


Fred: Editing Talk:info:gdfr/fred/f/pdf - Mozilla
http://tom.library.upenn.edu/cgi-bin/fred?cmd=EditWiki&sid=1b727c
Search
Home
Bookmarks
Editing Talk:info:gdfr/fred/f/pdf
This discussion page with your edits appears below. If further changes are needed, edit the form underneath. Use Fred's formatting conventions.
Your work will not be saved until you press the Save page button.
Thanks for updating this description, Stephen!
I notice that you've marked this as version 1.4. Since the name doesn't include that, is the assumption that this version will change to be whatever the "current" version is? (So that we'll update it to point to version 1.5 when that spec is out?)
-- John Ockerbloom
Thanks for updating this description, Stephen!
I notice that you've marked this as version 1.4. Since the name doesn't include that, is the assumption that this version will change to be whatever the "current" version is? (So that we'll update it to point to version 1.5 when that spec is out?)
-- John Ockerbloom

What's next for Fred?

- **Build a larger repertoire, community**
- **Elaborate on, discuss and evaluate data model details**
  - **Add support for classification systems, peer registries**
  - **How should classifications be designed and maintained?**
  - **Does data model need revision for reuse of agent and service information, peer exchange of information?**
- **Improve user interface, add machine interfaces**
- **Plan relationships between Fred and other registries**
  - **Including a successor GDFR system**
- **Explore synergies between Fred and TOM**
  - **TOM can bring in more sophistication about format relationships and abstraction, pointers to conversion services**
  - **Fred can bring better support for documentation, links to related description systems, support for version control and reflection**
- **Release code as open source**

What can you do to help?

See Fred's website: http://tom.library.upenn.edu/fred/

- **Register there to add and edit format information, participate in discussions**
- **Tell us what information, features are helpful to you, what changes you'd find useful**
  - **Practical, specific experience and applications especially helpful**