GDFR · Global · Digital · Format · Registry ·

Digital Library Federation Fall Forum
Boston, 8-10 November 2006

Global Digital Format Registry (GDFR)
An Interim Status Report

Stephen Abrams
Harvard University

Andreas Stanescu
OCLC

Global Digital Format Registry

- "The Global Digital Format Registry (GDFR) will provide sustainable services to collect, review, store, discover, and deliver significant representation information about digital formats."
  - Centrally-organized collection and review
  - Distributed storage, discovery, and delivery via a peer-to-peer network

Format and digital preservation

- Preservation is concerned with ensuring *access* to managed digital assets over time
- Thus, preservation activities are focused on
  - Viability
  - Fixity
  - Authenticity
  - Interpretability
  - Renderability
- The last two are primarily a function of *format*

Without format typing, all content is opaque

```
ffd8ffe000104a46494600010201
00830083000 0ffed0fb050686f74
6f73686f7020332e30003842494d
03e90a5072696e7420496e666f00
0000007800000000004800480000
000002f40240ffeeffee03060252
0347052803fc000200000480048
0000000002d80228000100000064
00000001000303030000001270f
000100010000000000000000000
000600800190190000000000000
000000000000000000000000000
000000000000000000003842
494d03ed0a5265736f6c7574696f
6e0000000010008313a3000200 ...
```

Without format typing, all content is opaque

```
ffd8ffe000104a46494600010201    SOI
008300830000ffed0fb050686f74    APP0  JFIF 1.2
6f73686f7020332e30003842494d    APP13 IPTC
03e90a5072696e7420496e666f00    APP2  ICC
0000007800000000004800480000    DQT
000002f40240ffeeffee03060252    SOF0  183x512
0347052803fc000200000480048     DRI
0000000002d80228000100000064    DHT
000000010003030300000001270f    SOS
000100010000000000000000000     ECS0
000060080019019000000000000     RST0
000000000000000000000000000     ECS1
000000000000000000000003842     RST1
494d03ed0a5265736f6c7574696f    ECS2
6e000000010008313a3000200 ...   ...
```

Without format typing, all content is opaque

```
ffd8ffe000104a46494600010201    SOI
008300830000ffed0fb050686f74    APP0  JFIF 1.2
6f73686f7020332e30003842494d    APP13 IPTC
03e90a5072696e7420496e666f00    APP2  ICC
0000007800000000004800480000    DQT
000002f40240ffeeffee03060252    SOF0  183x512
0347052803fc000200000480048     DRI
0000000002d80228000100000064    DHT
000000010003030300000001270f    SOS
000100010000000000000000000     ECS0
000060080019019000000000000     RST0
000000000000000000000000000     ECS1
000000000000000000000003842     RST1
494d03ed0a5265736f6c7574696f    ECS2
6e000000010008313a3000200 ...   ...
```

What is a format?

- Informally, “a serialized encoding of an abstract information model”
- Encompasses the nominal sense of “file format” as well as a range of conceptual entities from the micro to the macro level
  - IEEE 754 floating point number
  - File system

What is a format?

- Formal format model
  - AIM Abstract information model
    - FCS Format coded set (semantic)
    - FEF Format encoding form (syntactic)
    - FES Format encoding scheme (serialization)
  - SBS Serialized byte stream
- A format is a triple, *F* = (*FCS*,*FEF*,*FES*)
- Informed by the Unicode character encoding model www.unicode.org/unicode/reports/tr17/

GDFR project

- Two DLF-sponsored invitational workshops
  - University of Pennsylvania, January 2003
  - Washington, March 2003
- Provisional data and service models
- Two independent demonstration projects
  - FRED [John Ockerbloom, University of Pennsylvania] http://tom.library.upenn.edu/fred/
  - FOCUS [Joseph JaJa, University of Maryland] http://www.umiacs.umd.edu/~joseph/focus-archiving06.pdf

The GDFR project

- Harvard University Library (HUL) funded for 2 years by the Mellon Foundation
- Staffing and technical work subcontracted by HUL to OCLC (July 2006)
- Project oversight
  - Steering Committee (SC) for policy oversight
  - Technical Working Group (TWG) for technical oversight
  - Active solicitation of the international stakeholder community for review and comment

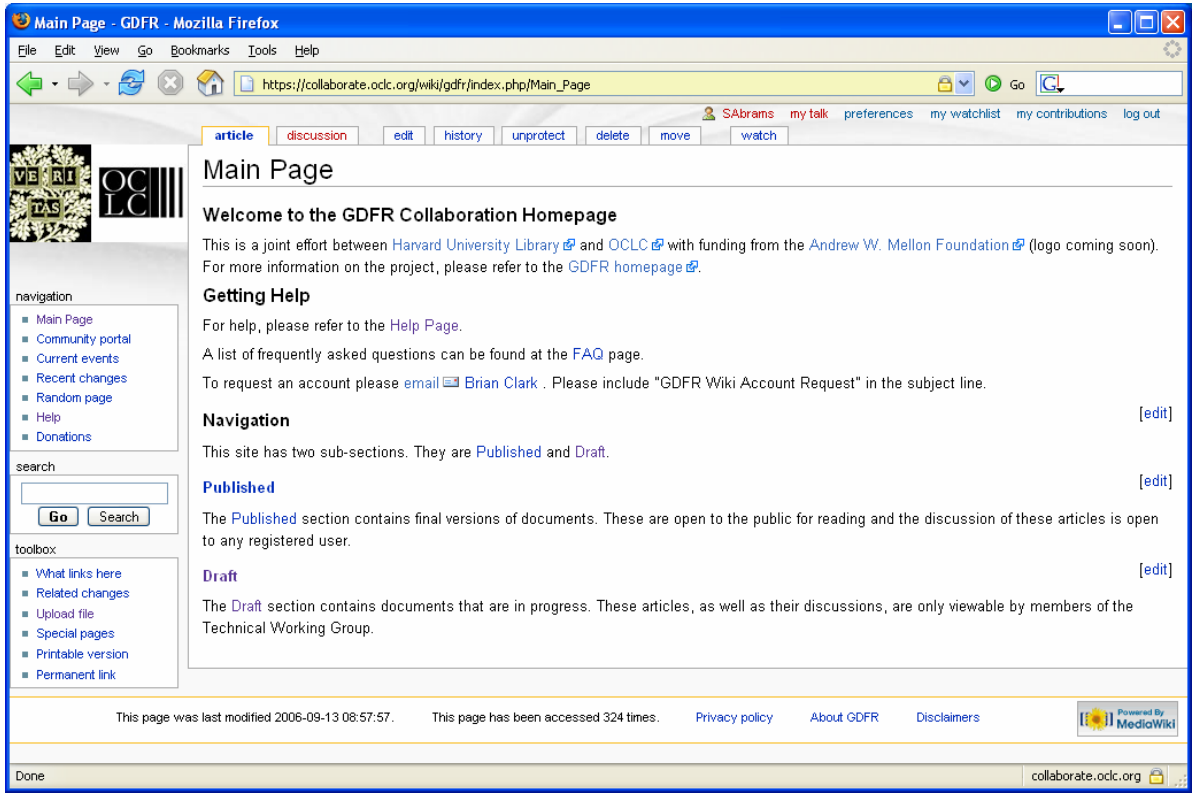
Technical Working Group (TWG)

- Bibliothèque nationale de France
- British Library
- California Digital Library
- Digital Curation Centre
- Library of Congress
- National Archives (UK)
- National Archives and Records Administration
- National Library of Australia
- National Library of New Zealand
- Stanford University
- University of Pennsylvania

Project web site and wiki

http://www.formatregistry.org

https://collaborate.oclc.org/wiki/gdfr/index.php



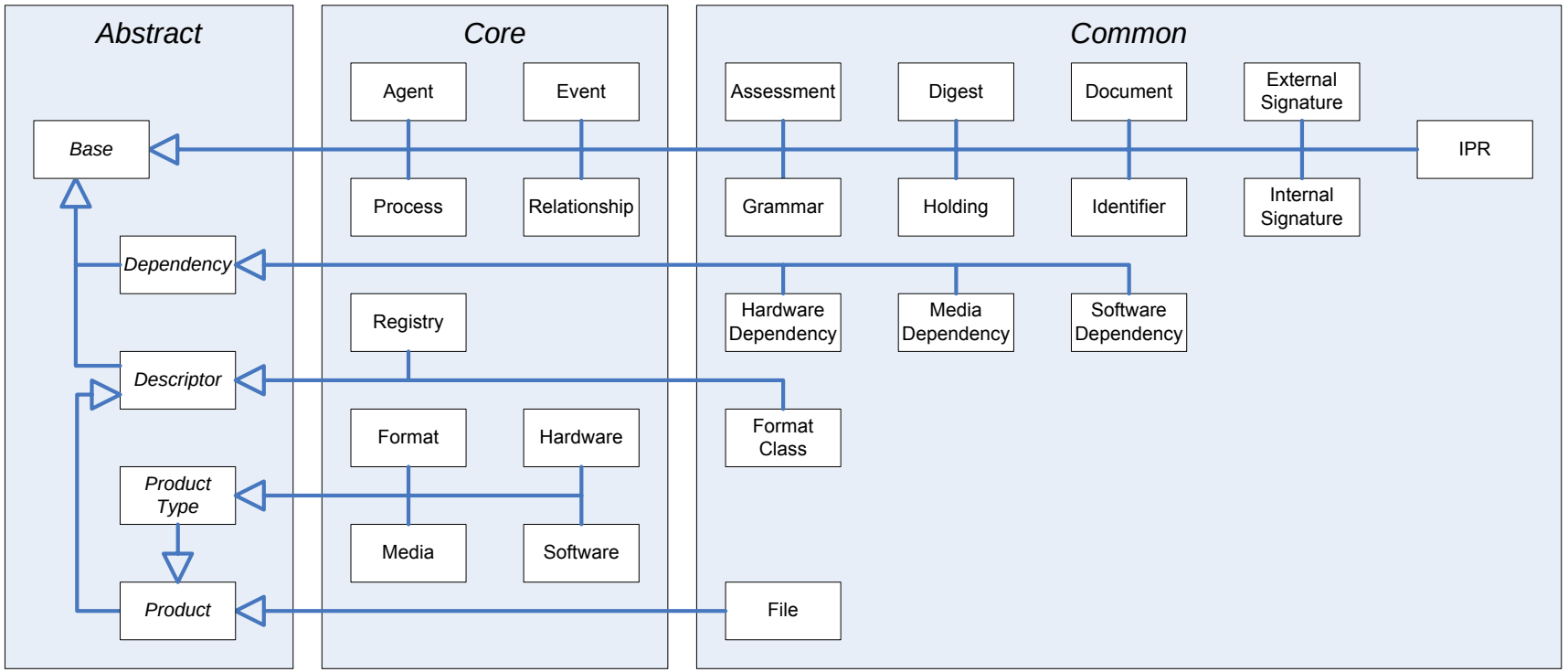
General development goals

- A generalized registry framework, specialized for the GDFR application
- Globally fault tolerant
- Platform independence
- Open source
- Re-use well-known products and protocols
- Human and machine interfaces
- Localization and accessibility
- Full information content expressible in XML form, and re-instantiatable from that expression

Data model

- ISO 11179, *Information technology – Metadata registries (MDR)*
- LC Digital Formats Web
  www.digitalpreservation.gov/formats/
- OASIS/ebXML Registry Information Model
  http://www.ebxml.org/specs/ebRIM.pdf
- PRONOM
  www.nationalarchives.gov.uk/pronom/
- Representation Information Registry/Repository
  dev.dcc.ac.uk/twiki/bin/view/Main/DCCRegRepV04

Data model


Abstract
Core
Common
Base
Dependency
Descriptor
Product Type
Product
Agent
Event
Process
Relationship
Registry
Format
Hardware
Media
Software
Assessment
Digest
Document
External Signature
Grammar
Holding
Identifier
Internal Signature
IPR
Hardware Dependency
Media Dependency
Software Dependency
Format Class
File

Format entity

- Canonical (GDFR) and alias identifiers
- Version
- Description
- Classification
- Relationships
- Disclosure – open, proprietary, closed
- Documentation
- Orientation – text vs. binary
- Byte order

Taxonomy

Ontological CLASSES, abstract *families*, concrete formats, and **relationships**

BYTESTREAM			
IMAGE			
STILL			
RASTER			
GIF			
	GIF87a		
	GIF89a	**is-new-version-of**	GIF87a
JPEG			
	ISO 10918-1		
	JFIF	**is-extension-of**	ISO 10918-1
TIFF			
	TIFF 4.0		
	TIFF 5.0	**is-new-version-of**	TIFF 4.0
	TIFF 6.0	**is-new-version-of**	TIFF 5.0
	TIFF/IT	**is-extension-of**	TIFF 6.0
	TIFF/IT/CT	**is-subtype-of**	TIFF/IT
	TIFF/IT/CT/P1	**is-subtype-of**	TIFF/IT/CT

Relationships

• Subtype	ASCII	**is-subtype-of**	UTF-8
• Extension	DNG	**is-extension-of**	TIFF 6.0
• Containment	WAVE	**can-contain**	μ-law
• Equivalence	DXF(ASCII)	**is-equivalent-to**	DXF(binary)
• Version	TIFF 6.0	**is-version-of**	TIFF 5.0
• Affinity	SPIFF	**is-similar-to**	JPEG

Documentation

- Public domain specifications managed and replicated in the network
- For non-public domain, full bibliographic citation with actionable identifiers
- Mechanism for agents to register locally-held copy with terms of use

Format entity

- Internal/external signatures – magic number/file extension
- Grammar – ABNF, BNF, BSDL, DFDL, EAST
- Assessment – LC SQF, OCLC INFORM, DSTC PANIC, VRC
- Release date
- Withdrawal date
- Software, hardware, media dependencies
- Developer
- Support
- Rights

Domain model

- Actors
  - Public user
  - Registry node
  - Registry editor
  - Registry administrator
  - Registry superuser
- Use cases
  - Generic registry
  - Distributed registry
  - GDFR

Use cases

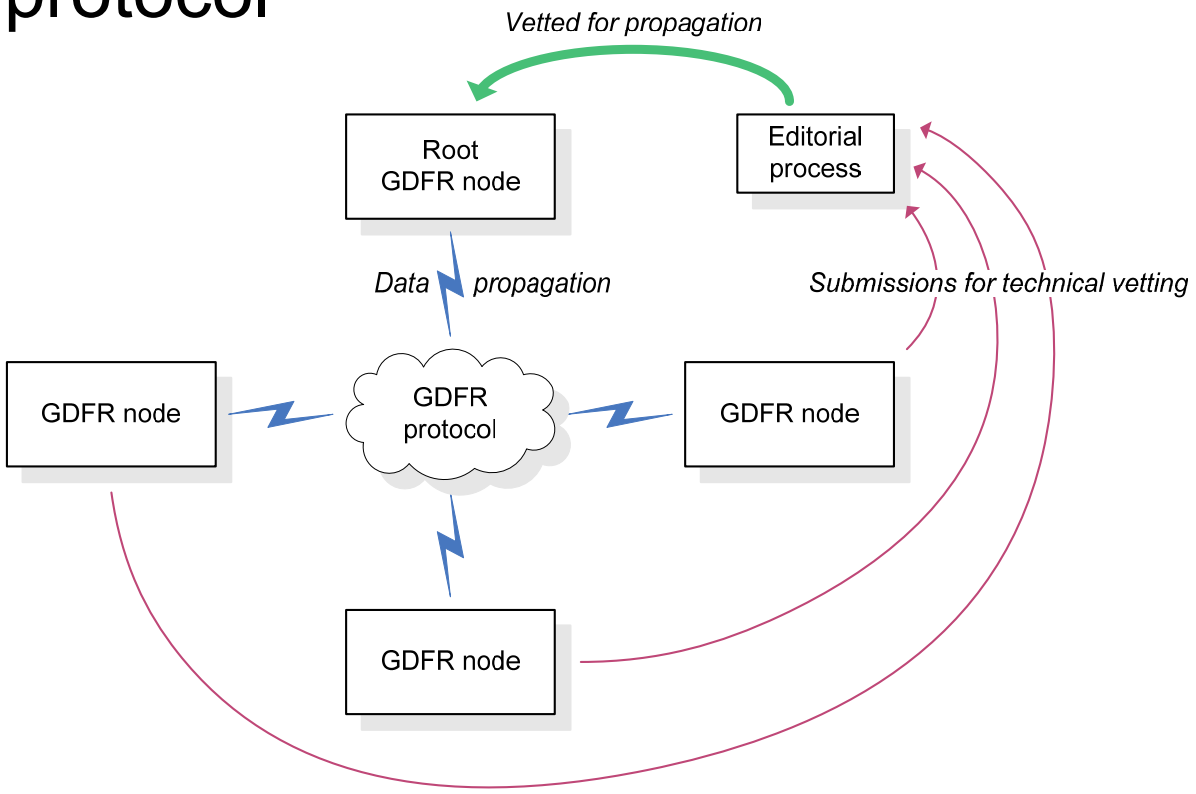
- Discover registry
- Authenticate user
- Recognize user
- Search collection
- Read record
- Export records
- Import records
- Add collection
- Configure collection
- Create record
- Add record
- Update record
- Purge record
- Recognize peer registry
- Distribute records
- Synchronize records

Service model

- Maintenance
  - Add, review, update, and store representation information
- Patron
  - Manual and automated discovery and delivery of representation information
- Local administrative
  - Policy
- Global administrative
  - Distribution
  - Synchronization

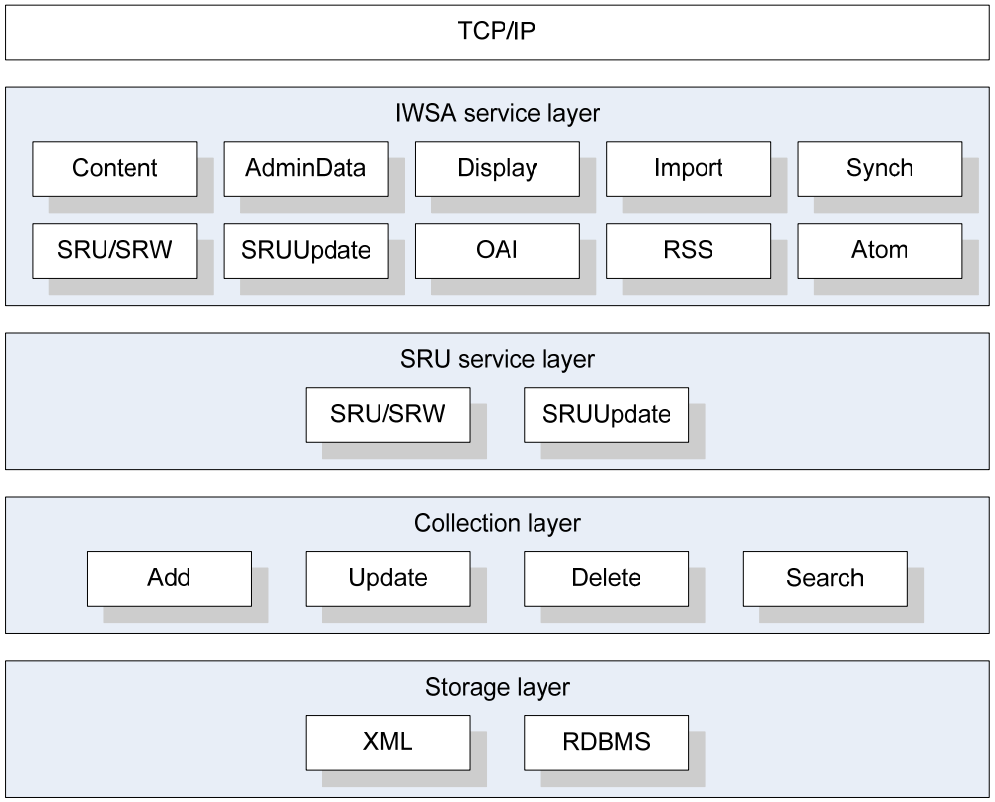
GDFR network

- Peer-to-peer network communicating over a common protocol



GDFR node

- Based on the IWSA framework



Technical review

- RFC 2026, *Internet Standards Process*
  www.ietf.org/rfc/rfc2026.txt
  - "Iterations of review by the ... community and revision based upon experience"
- Draft distribution and public discussion
- Approval by "area" editors
- Release to the network for distribution

Governance and succession

At the end of the two year Mellon-funded project, GDFR will be turned over to a long-term policy and maintenance agency

- Harvard will undertake to continue maintenance for up to two years
- Library of Congress has agreed to be a care-taker agency until a permanent body is identified

- NARA GDFR governance investigation

Summary

- The GDFR is an enabling technology that will support digital repository operations and preservation activities
  - Enables the typing of digital objects at an appropriate level of granularity
  - Enables the future recovery of the syntax and semantics associated with typed digital objects
  - A means to pool and redistribute the expertise of the digital preservation community

For more information

Birds-of-a-Feather (BOF) session
(for background, see http://hul.harvard.edu/gdfr/documents.html)

http://www.formatregistry.org/
https://collaborate.oclc.org/wiki/gdfr/index.php

stephen_abrams@harvard.edu
stanesca@oclc.org