Linked People

Building a community around trustworthy data

Andrew Jackson

The British Library

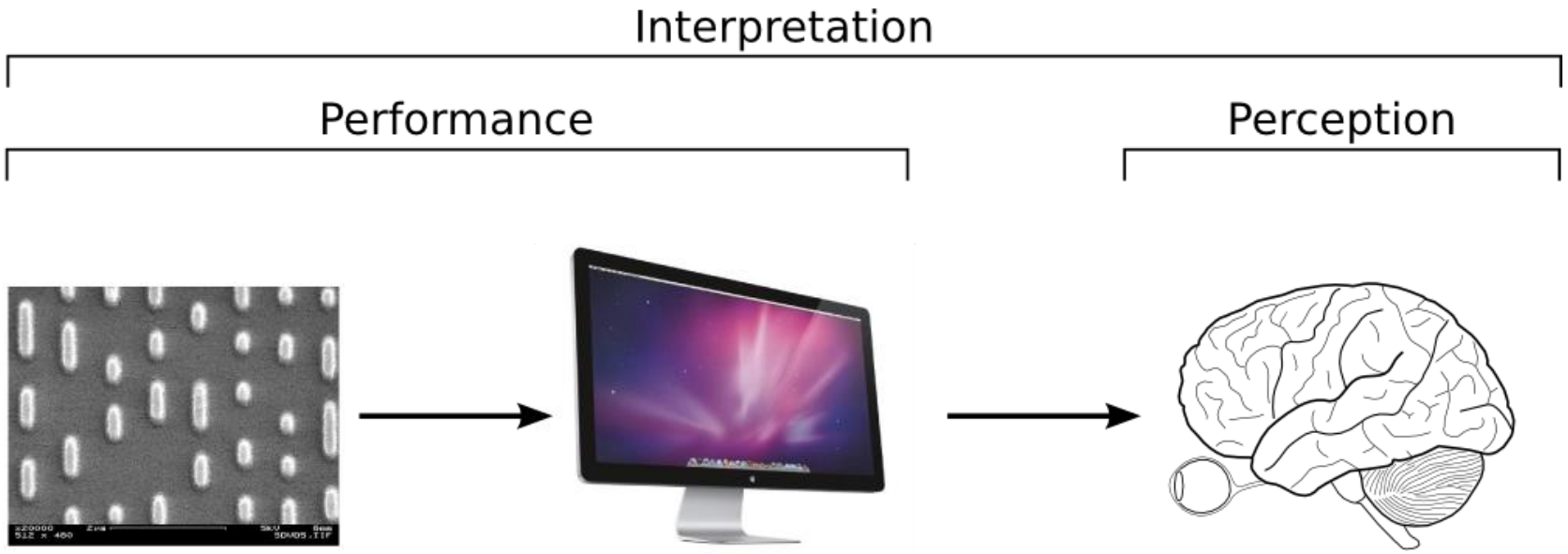
The Future of File Format Identification: PRONOM and DROID User Consultation
The National Archives, Kew, London, 28.11.2011

Identification Tools

- DigiPres Community:
  - DROID
    - PRONOM
    - Bitstream + Container ID
  - FIDO
    - PRONOM (via RegEx)
    - Bitstream ID only
  - JHOVE
    - Identify by parse
  - NZ Metadata Extractor
    - Adaptors and Recognisers
- Rest Of The World:
  - The File command
    - Bitstream ID and some Container ID support
  - Apache Tika
    - Bitstream, Container and XML namespace ID
  - File Investigator Tools
    - Bitstream ID, >4100 formats supported
    - Commercial ($249)

It's Not Just About Identification

- Preserving representation Information (RI):
  - The information required to interpret a digital object



DVD image from http://www.optics.rochester.edu/workgroups/cml/opt307/spr05/chris/

Preservation Registries With Data

- PRONOM & Linked-Data PRONOM
  - http://www.nationalarchives.gov.uk/pronom/
- Library of Congress Format Registry
  - http://www.digitalpreservation.gov/formats/
- The Software Ontology
  - http://theswo.sourceforge.net/
  - Software focused, treats format as class
- KEEP TOTEM
  - http://keep-totem.co.uk/
  - Model covering hardware as well as software and formats

KEEP TOTEM PDF Summary

KEEP TOTEM PDF Versions

PUID: fmt/19

Version Name: PDF 1.6

Related To:

Previous Version Of: PDF 1.7

PUID: fmt/20

Version Name: PDF 1.7

Related To: ISO 32000-1:2008

Previous Version Of: PDF 1.7 Extension Level 3

PUID: fmt/276

Version Name: PDF 1.7 Extension Level 3

Related To:

Previous Version Of: PDF 1.7 Extension Level 5

PUID:

Version Name: PDF 1.7 Extension Level 5

Related To:

Previous Version Of: PDF 1.7 Extension Level 8

PUID:

Version Name: PDF 1.7 Extension Level 8

Related To:

Current Solutions

- Monolithic Registry Solutions:
  - GDFR & UDFR
    - http://www.gdfr.info/ & http://www.udfr.org/
  - The CASPAR/DCC RI Repository
    - http://registry.dcc.ac.uk:8080/RegistryWeb/Registry/
  - Complex architectures built in isolation
    - Contain little or no data
- All our registries are desert islands
  - Thinly populated
  - Poorly linked

The Rest Of The World's Registries

- MIME Media Types
  - http://www.iana.org/assignments/media-types/
- Wikipedia
  - http://en.wikipedia.org/wiki/Category:Computer_file_formats
- Freebase
  - http://www.freebase.com/view/computer/file_format
- W3C's Ontology for Media Resources
  - http://www.w3.org/TR/mediaont-10/

Registry Eco-System

- Working Group to create:
  - Initial data model
  - Guidelines for publishing the format data
- Even more sources of data:
  - How will the data be consumed?
  - How will we know we can trust the data?
  - Trust needs more than provenance
  - How do we build trust in the data itself?

Raising The Quality

- Quality assurance
  - Community data quality guidelines
  - Aggregated & evaluate:
    - Automated testing
    - Open peer review
  - Publish quality metrics & feedback
- Sustainability
  - Integrated part of user tools & workflows
  - Growing the community
  - The promise of persistence

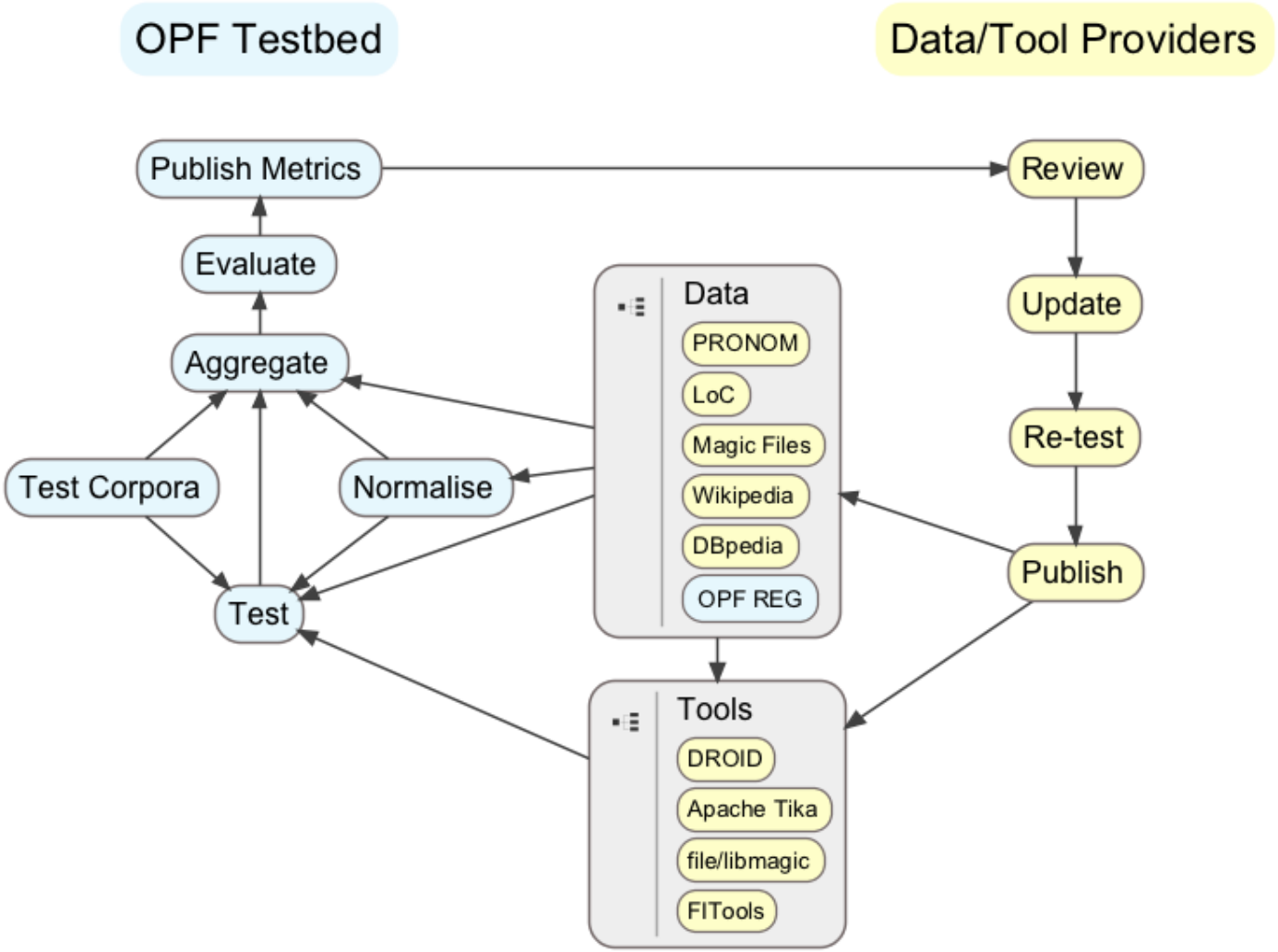
Quality Assurance: Data Quality Guidelines

- Required, desirable and optional fields
  - e.g. Each format description:
    - MUST have a name, a link to formal standard and/or reference implementation, and at least one sample file
    - SHOULD have a version, extension, MIME Type, etc.
- Keep facts separated from policy
- Every factual field should be verifiable
  - e.g. There should be test files to run signatures against
  - e.g. MIME Type field linked to IANA registry
- WG to develop initial guidelines?

Quality Assurance: Testing The Data

- Centralised assessment of registry data
  - Just enough centralisation to bring some continuity
  - Provide a focus point to encourage coordination
  - Provide tools to help verify the data
- Consume linked data or other data sources
  - Normalise to agreed format if necessary
- Expose the conflicts and encourage resolution
  - Compare data against guidelines
  - Compare data sources against each-other
  - Look for gaps and inconsistencies

Public QA Cycle



OPF Results Evaluation Framework

Login Create account

About | Members | Events | Projects | Community | Contact

http://data.openplanetsfoundation.org/ref/data/ooo/ooo

Alternative Formats: RDF

opf-ref:has-result	http://data.openplanetsfoundation.org/ref/results/27531
opf-ref:has-result	http://data.openplanetsfoundation.org/ref/results/26549
opf-ref:has-result	http://data.openplanetsfoundation.org/ref/results/18718
opf-ref:has-result	http://data.openplanetsfoundation.org/ref/results/36877
opf-ref:has-result	http://data.openplanetsfoundation.org/ref/results/16512
opf-ref:has-result	http://data.openplanetsfoundation.org/ref/results/17738
opf-ref:has-result	http://data.openplanetsfoundation.org/ref/results/20627

Files/Tools	0.9.3-50 fido	6.0.1-50 Droid	4.0-50 Droid	2.32 File Investigator	5.04 file	4.0-16 Droid	0.9.5-50 fido
	OLE2 Compound Document Format fmt/111	Microsoft Powerpoint Presentation 97-2002 fmt/126	fmt/126 Microsoft Powerpoint Presentation		CDF V2 Document, Little Endian	fmt/126 Microsoft Powerpoint Presentation	OLE2 Compound Document Format fmt/111

Sustainability: Integrated Into User Workflows

- Understand user workflows
  - How is Droid is used? Or are other tools in use?
  - Make the benefits clear
    - file/FITools plus Wikipedia and you are 95% done!
  - Make adding information easier
    - Make local extensions easy to add and to share
    - Lower the technical barrier via common sig. formats, e.g. RegEx
    - Use DROID data with file/Tika engine? (More users, less code!)
- Encourage a culture of data sharing
  - Publishing data as business as usual
    - 'Publish' buttons/switches in users' tools?

Sustainability: The Promise Of Persistence

- Memory Institutions can really help by providing
  - Support
    - Some effort and/or money needed for WG/Data QA system
    - Public approval of process and data quality guidelines
  - Goals
    - Quality and coverage goals that trigger ingest, or even prizes?
  - Persistence
    - Submit high quality data to PRONOM for consideration
    - Releases archived in memory institutions
- Build new relationships:
  - Wikipedia and other/future preservation projects

Conclusion

- We need to find a way to work together
  - Grow the eco-system of registry data
  - Add just enough centralisation to:
    - Perform QA, provide feedback
    - Encourage convergence & sustainability
    - Grow the community: link up the people as well as the data
- This meeting is a great opportunity
  - But we need simple tools and dedicated resources
  - Share the data and let it lead the way

Quality Assurance

Quality Assurance: Aggregate & Evaluate

- Consume linked data or other data sources
  - Normalise to open standards
- Expose the conflicts and encourage resolution
  - Compare data against guidelines
  - Compare data sources against each-other
  - Look for gaps
- e.g. Internal File Signatures
  - freedesktop.org Shared MIME Info Specification
  - plus interoperable identifiers:
    - ‘application/pdf; version=1.4’ === ‘info:pronom/fmt/18’

freedesktop.org Shared MIME Info Specification

```
    <glob pattern='*.onetoc2'/>
    <glob pattern='*.onetmp'/>
    <glob pattern='*.onepkg'/>
  </mime-type>
  <mime-type type='application/parityfec'/>
  <mime-type type='application/patch-ops-error+xml'>
    <glob pattern='*.xer'/>
  </mime-type>

  <mime-type type='application/pdf'>
    <alias type='application/x-pdf'/>
    <acronym>PDF</acronym>
    <_comment>Portable Document Format</_comment>
    <magic priority='50'>
      <match value='%PDF-' type='string' offset='0'/>
    </magic>
    <glob pattern='*.pdf'/>
  </mime-type>

  <mime-type type='application/pdf; version=1.4'>
    <sub-class-of type='application/pdf'/>
    <acronym>PDF 1.4</acronym>
    <_comment>Portable Document Format - Version 1.4</_comment>
    <magic priority='70'>
      <match value='%PDF-1.4' type='string' offset='0'/>
    </magic>
    <glob pattern='*.pdf'/>
  </mime-type>

  <mime-type type='application/pgp-encrypted'>
    <glob pattern='*.pgp'/>
  </mime-type>
  <mime-type type='application/pgp-keys'/>
  <mime-type type='application/pgp-signature'>
```

Quality Assurance: Automated Testing

- Testing format data automatically
  - A ‘lint’ tool to check data meet quality guidelines
  - Report includes information on how to improve
  - Runs centrally and can be used locally
- Testing identification tools automatically
  - Functional testing for preservation identification
    - Run ID tools on a large corpus, look for gaps & disagreements
- Compare all results via a shared database
  - Results Evaluation Framework
    - http://data.openplanetsfoundation.org/ref/data/000/000725.ppt

Quality Assurance: Open Peer Review

- For the things we cannot test automatically
- Must be done in the open
  - Having one-to-one conversation doesn't scale
  - Published communications and discussions help others to understand the issues and learn how to contribute
  - Major stakeholders should play active roles
- Applies to the schema and the quality assurance framework as well as the data itself

Quality Assurance: Quality Metrics

- Per-entry metrics
  - Peer review status
  - Number of test files
- Aggregate metrics
  - ‘Fullness’ metrics
    - Total number of entries
    - Percentage of entries that meet data quality guidelines
  - Quality metrics
    - Percentage that have been peer reviewed
    - Percentage that have at least one test file
    - Percentage of test corpus covered

Modelling

Building Our Data Model

- Open data model, as simple as possible
  - Working Group set up to create initial model
- Mind The Scope
  - Needs Driven
    - When is RI used?
    - What are the common business processes?
  - Data Driven
    - What events herald the birth of a new format?
    - How do formats ‘die’?
    - What are the difficult identification cases?
- Expect the schema and the data to grow organically
- Linked Data/RDF model, hosting and discoverability

It's Not Just About Filling Registries

- We're not going to hire people to fill out registries.
- We do normal work, but want to share results to make things easier.
- A full and active RI registry would reflect a community that understand data formats and wants to share that understanding.
- **That community of expertise is what we need!**
  - Tools and technical registries are reflections of those minds, and filling them is a side-effect of the work those people do.

RI Is Infinite

- Which RI do we need?
- ** Add new records, have quality standards, submit to PRONOM for consideration.
- ** Collect gaps, e.g. formats that don't fit or need more fields etc.
- ** Evolve the standards based on this data and gaps.
- ** Use source code control techniques to merge and synchronise different data sources.

Modelling Format Is Hard

- RI & Persistence
  - Record mixed spec. software and other stuff.
- * Modular Design
- ** Identification signatures cleanly separated from RI, but linked via identifiers.
- ** Similarly, keep new concepts apart but linked via URIs. e.g. KEEP could link to format identifiers.

Modelling Is Endless

- Model, treats spec as king.
  - But Software is king.
  - Mutants and wild types. Strains. Quirks Mode.
- Modelling will be hard, so we must be fluid.
- Let the data show the way.
- The Schema Will Change

The Anvil

- * The Anvil
- Corpora based testing, covering all the most difficult cases we can think of.
- Test identification tools and registries for coverage and consistency.
- Also need test structures of documentation, peer review?

Engage With The Broader Community

- We are not alone
- Ref SCAPE Work Johan tools existing.
- Tika

The Problem

- All these tools, all these registries
  - Massive reproduction of effort
  - Why are they isolated? Why are they bare?
- How do we start bringing these threads together?
  - How to share data and build something together?
  - How to agree a way of talking about formats?
  - How to make it easy to share data?
  - How to ensure contributors feel valued?
  - How to ensure the information is trustworthy?
  - How to engage with the wider community?