Towards a Format Registry Ecosystem

Bill Roberts

National Archives of the Netherlands

‘Ecosystem’?

- Info on formats/software from lots of places
- Means of exchanging it
- Means of adding your own or enriching other sources
- Means for users to select, gather, manage

Possible models

One ring to rule them all

Let a hundred flowers blossom

‘encyclopedia brittanica’ vs ‘wikipedia’ vs ‘the web’

PRONOM

TOTEM

NLNZ ‘Technical Library’

SCAPE

UDFR

KEEP

Freiburg emulation environments

Wikipedia

Commercial tools

Open Planets Foundation

Hamburg workshop

- Yes we need format registries
- Core is identifiers and corresponding ‘definitions’
- But want as comprehensive a description as possible (also ‘policy’/usage/opinion as well as fact)
- Need to share the work
- Provenance info is essential
- Need to agree ‘guidelines’: data model, exchange format and common approach to identifiers
- Establish a working group

Working group

- No face to face meetings
- Results by end Feb 2012
- Open to anyone – but want active contributors
- Data model
- Exchange format (will be RDF – will look at ontologies)
- Approach to identifiers
- Want to be involved? See me after…

What do we need to think about?

- Complexity of information
- What are we identifying?
- Making exchange and re-use easy
- Helping people use the data
- Trust and provenance

realists

Registries and preservation for ~~beginners~~

- What am I going to do with file F?
- F has format X
- Format X can be read by software S
- I've got a working copy of S
- ☺

What do you mean by X?

Well, mostly…but the fonts are wrong

Are you sure?

What are the dependencies of S?

Open Planets Foundation

Identifiers: What are we identifying?

- PRONOM vs Outside In vs File Investigator
- New registries minting own identifiers
- File format defined by:
  - A specification ?
  - Match to a signature?
  - Whether it can be read by a particular tool?
- How to interlink different identifiers? (owl:sameAs, ‘kind of the same as’, ‘superset of’…)

Technical requirements

- Information needs to be precise – to support sensible decisions
- Machine readable – to support automation
- Independent of any particular software system – standards-based, exchangeable, reusable, preservable
- Easily extensible to support specialist requirements

Using the data

- Tools for working with multiple sources of representation information?
- ‘Format dashboard’
- Vendor support?

Trust and provenance

- How to decide which data sources to use?
- Who published the information?
- How did they produce it?
- What process of checking and testing have they done?
- Do other people trust it and use it?
- If I test it, how do I share that information with others?

Challenges for registry guidelines

- Balancing flexibility and interoperability
- Ensuring compatibility with PRONOM but not being limited by it.
- Making the guidelines ‘as simple as possible but no simpler’.
- Getting people to support them and use them:
  - Helping content-owners get started
  - Helping data consumers get started