Metadata, preservation, and sustainability

Simon Mahony

From an original document by Claire Warwick

Version 1.0

The Problem

- Long term sustainability of digital data
- Digital medium unstable
  - Paper still preservation medium
- How to ensure trustworthiness of digital resources
- How to organise the web
- How to find things on the web
- Information about resources needed

The semantic web

- •Web is inherently disorganised
- •Search engines can only index small portion
- •Tim Berners-Lee suggests Semantic web
  - –Use of XML to help add information to pages
  - –Sometimes called web 3.0 (or even 4.0)
- •Metadata essential to this process
- •Use of Ontologies (more on this later)
  - –Cataloguing and Classification for the web

Metadata

- Data about data
  - Like catalogue records for books
- Essential for semantic web (linked data)
- Also vital for usability and sustainability of materials
- Possible to apply this to HTML pages using <meta> tags in the <head>
- Not seen by human readers
- Problem with early abuse
  - Not widely used by search engines

Metadata schemes

•Community specific

–TEI header, very detailed information in the header of the XML file

–Highly structured and directive

•Dublin Core

–Simple set of 15 elements

–Less direction about use

•Now integrated within RDF Schema for use with Semantic Web (Resource Description Framework)

•METS: Metadata Encoding & Transmission Standard

–XML schema used in digital libraries

–MARC (standard): Machine Readable Cataloguing

Dublin Core elements

- Title
- Creator
- Subject
- Description
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights

The problem with metadata

- Has historically been an expert activity

– Librarians catalogue and classify things

- Element sets either too complex or not easy to understand

- Who is the creator: the author, publisher, library?

- Non professionals don't see the need for it

- So much content now created by amateurs

- But without metadata it's hard (or impossible) to find things

Sustainability

- Digital resources being lost
  - Especially in non-commercial sector
  - But not exclusively – Doomsday Project (laserdisc)
- Losing functionality
- Hardware and software issues
- Lack of maintenance
- Lack of updating
- Funding for doing this missing
  - But complete loss of resource greater cost

Why does this matter?

- Ever larger amounts of digital data being created
- Little expertise about how to preserve it
  - What to preserve
  - What to weed out?
- Archivists know about this
  - But it's often being done by systems managers
- Institutional repositories not sure how to treat data
- What to accept, in what format?

Designing for Sustainability

- Use of open standards and open source in software
  - Platform specific software problematic in the long term
- Need for migration and emulation
- Informed decisions about what to keep
- Greater awareness of changeable nature of web-based resources
- Need for metadata and documentation

Metadata and Sustainability

- Metadata needs to be accompanied by documentation
- Together provide information about rationale for resource creation (why has it been done in the way that it has?)
  - User of materials and techniques
- More information provided, easier migration or reconstruction may be
- Helps to preserve memory of long term projects
  - Institutional knowledge management vital

Metadata and Users

- Users require information about digital resources
  - Especially academic or expert users
- Selection of content
- Extent of resource
  - If selected what methods used
  - If updated how often
- Who is responsible for resource?
  - creation, publishing, maintenance
  - needs to be built into the design and workflow

Metadata and Users 2

- What sources have been used to create the resource
  - If data created then by whom?
- How reliable is the 'publisher'?
- Reliability of resource long term
- Where help is available if necessary

Current situation

- Metadata and documentation rare, especially in the case of non-commercial resources
- May be difficult to find if existent
- Users may not make it to the metadata if other things put them off
  - Thus must not be relied upon to make up for other deficiencies
- Increases trust in resource if present

Prospects for future

- How to make the semantic web a reality?
  - Why should people add metadata to resources?
  - Google seems to work OK
  - Commercial costs of doing this
- Is it inevitable in the age of Google Book Search and e-science?
- Watch this space!

Problems for web publishers

- Packaging information
  - selling ways of thinking about content
- Expectation of currency can be a burden
- Maintenance and support