The Metadata Challenge: Determining local and global needs and expectations for your metadata

Kultivate Metadata workshop
24th May 2011

Gareth Knight,
Centre for e-Research (CeRch), King’s College London
Presentation structure

1. Role of metadata in a digital repository
   1. How do you determine fitness for purpose?

2. User 1: Local researchers
   1. What do they wish to achieve? What do they require?
   2. Do you currently fit their needs?
   3. How can you better address needs?

3. User 2: 3rd party harvesters
   1. What questions may they have?
   2. What do they require? Balancing local & global needs

4. User 3: Staff involved with REF preparation
   1. What do they wish to achieve? What questions may they have?
   2. Approaches to addressing questions – work performed in Readiness 4 Ref project

5. Conclusions
Metadata and its evolving role

“metadata is developed by people for a purpose or a function”

The traditional library

The digital library

metadata

…and in the wider world
Is your metadata fit for purpose?

1. **Who and where are the stakeholders? Who are the audience who will use your metadata?**
   - People: staff, students, researchers, etc.
   - Machine: Repository registries

2. **What do the users wish to achieve when using the metadata? What function does it need to perform?**
   - Do they have specific questions that they wish to address.

3. **How will your metadata fulfil these functions?**
   - Implications for quality assurance: Is it good enough for the intended purpose?
   - In what format do they expect a response to be provided?

4. **If do not currently meet their needs, how do you intend to provide them with the information they require?**
   - Metadata creation - implications of time / resources
   - Machine vs. user creation
What are the stakeholders of a Creative Arts repository?

- Researchers of different types at different levels
  - Artists,
  - Performers,
  - Undergraduate and postgraduate students
- Staff preparing for REF submission,
- Funding agencies who wish to access research
- 3rd party subject repositories and value-added services that wish to obtain and use your metadata
- ... and others
User 1: What do researchers expect from your metadata?

Discovery
• What resources exist on topic?
• What resources were created by this person?

Identification
• What is it?
• Who created it?
• What does it represent? What does it contain?
• When was the physical original created?
• When was digital edition created?

Use:
• Can I use it in my own work?
• How do I cite it?

Is your metadata fit for purpose? Does it answer researcher questions?
# Adolphe Appia collection metadata

<table>
<thead>
<tr>
<th>Title</th>
<th>Adolphe Appia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Adolphe Appia, the Swiss theatrical practitioner who died in 1928, revolutionary stage practice. Appia's imagination went far beyond his groundbreaking in space. His theories and practical investigations set an agenda for the use of images and virtual reality models provide vivid and varied demonstration</td>
</tr>
<tr>
<td>Format</td>
<td>Images; VRML models</td>
</tr>
<tr>
<td>Extent</td>
<td>230 images</td>
</tr>
<tr>
<td>Date(s)</td>
<td>(Created): 01 Jan 2004</td>
</tr>
<tr>
<td></td>
<td>(Deposited): 01 Jan 2005</td>
</tr>
<tr>
<td>Subject</td>
<td>(AHDS terms): Drama and theatre</td>
</tr>
<tr>
<td></td>
<td>(LCSH terms): Theatrical production and direction; Theatrical producers and managers</td>
</tr>
<tr>
<td>Administrator</td>
<td>AHDS Performing Arts</td>
</tr>
<tr>
<td>Availability</td>
<td>Available for non-commercial, educational use only</td>
</tr>
<tr>
<td>Rights</td>
<td>Copyright King's Visualisation Lab, King's College London</td>
</tr>
<tr>
<td>Status</td>
<td>This collection is closed to further updates/deposits. Possibly. Contact details?</td>
</tr>
</tbody>
</table>

Q1. What is it?
Q2. Who created it?
Q3. What does it contain?
Q4. When was the physical original created?
Q5. When was digital edition created?
Q6. Can I use it in my own work?
Q7. How do I cite it?
Q8. What other resources exist on topic?
Q9. What other resources were created by the author?

Answers some, but not all questions
Implications

Metadata answers some, but not all questions

Questions:

1. How can we make better use of existing metadata?
   - Perform NLP to extract information from large body of text
     - Reformat data elements into consistent format, e.g. creator names
     - Map depositor free text md to structured subject terms

2. What additional metadata is required to fit evolving needs?
   - How will metadata be extended and enhanced?
     - Do you have the resources to enhance catalogue metadata
User 2. What do 3rd parties expect?

A 3rd party wishing to harvest your MD may have several questions:

1. How do I identify the MD records that repository X has published?  
   OAI-PMH, RSS, OAI-ORE

2. Can I obtain all metadata for an item?  
   Must determine exposed formats – OAI-PMH

3. Are we allowed to reuse and republish the records (rights)?  
   Anyone have a machine-processable MD policy?

4. Is it possible to combine MD records from multiple repositories to provide value added features?  
   Reuse of MD can prove to be a challenge
# Metadata 101

Metadata is composed of three components:

1. **Semantics**
   - The meanings of the metadata elements

2. **Content rules**
   - What values go into the elements

3. **Syntax**
   - How they are encoded

## Metadata Elements

| Title       | Mandato
|-------------|--------|
| **DC**      | The name given to the resource by William Shakespeare
| **MODS**    | Optional

Metadata standards address and resolve these issues, but only if used correctly.
Interoperability challenge: Balancing local & global needs (1)

Scenario:
Repository have local requirements and wish to catalogue information not found in existing schema (e.g. death date). Shoehorn information into existing element, Claim that local implementation conforms to standard

Potential implications:
• Worse case scenario - MD originating from multiple sources and combined into single scheme will contain information that has different semantic meaning and content rules – user may be misled by information
• Each repository has different control rules, resulting in variation in vocabulary restrictions, differences in handling of names, subject terms, rights
  • AHDS, SHERPA DP preservation repository examples.
• Some elements may be omitted when converting between metadata formats using standard crosswalks
  (e.g. http://www.getty.edu/research/conducting_research/standards/intrometadata/crosswalks.html)
Interoperability challenge: Balancing local & global needs (2)

Possible Solutions:

• Application profile
  • Declare implementation is an (DC, VRA, CDWA, MODS, etc.) application profile with some local elements

• Packaging format, incorporating multiple schemas
  • Utilise MD packaging format (METS, MPEG21) and combine two schemes
New and emerging purpose/function:

Research Excellence Framework

Research support staff preparing information for REF submission may be considered a new type of stakeholder that must be considered when creating & enhancing metadata
Potential approaches

1. Repository plus in-house research database (KCL approach)

2. Repository and CRIS in concert

3. Standalone repository
Research Information systems at King’s

- **HRMS**: Name, job title, start date, etc. Feed automatic, overnight
- **SITS**: PG(R) details. Feed automatic, overnight
- **APTOS**: Grant monthly expenditure. Feed automatic, overnight
- **Res. Grants & Contracts**: Awards, amounts, dates. Feed manual, 2 weeks
- **IoP Res. Grants**: Awards, amounts, dates. Feed manual, 1 month
- **Web of Knowledge**: Publications in journals. Feed automatic, 1 week
- **RG Academic Repository**: Full text of publications, publication metadata
- **School/College Managers – Research Strategy**: Direct link
- **School web pages (potential)**
- **Experts database (potential)**
- **Staff Profiles (Web pages)**
- **RAE/REF outputs**
- **AKORD publications database (web interface)**
- **InCites - bibliometric information**
Why use a digital repository to create metadata for REF?

Benefits:
- Established infrastructure - core part of institutions, trusted publisher
- Accustomed to handling information flows – OAI-PMH, SWORD, RSS
- Sustainable over time – not just one-off submissions to REF

Challenges:
- Many institutions do not have all information required for REF
- Repository metadata insufficient for REF requirements
  - Many repository data models are too simple for REF
    - REF entities require more than just a data element in an ePrints object record
    - VRA, CDWA, MODS, etc. do not support element set required by REF
User 3. What do staff involved with REF preparation expect?

1. Where do I locate information on research structures, staff, output and other entities?
2. Can I identify and obtain all metadata of relevance from the repository (in order to analyse and merge it at a later date)? (Similar to user 2)
3. Is metadata provided by digital repository (and other systems) of sufficient quality and completeness for REF submission?
4. Is it possible to relate research artefacts referenced in IR to metadata stored elsewhere?
5. Can I ensure that REF metadata is stored and exported in manner consistent with other UK research institutions?
Readiness 4 Ref

Readiness 4 Ref (R4R) funded to address some of these questions.

1. Survey existing university systems
2. Analyse REF requirements
3. Apply CERIF to REF
   - CERIF provides many of the descriptive elements for capturing REF requirements.
   - Devise CERIF4REF XML application profile
4. Case studies in HEIs, funders, publishers
   - How easy is it for institutions to provide their data using the CERIF4REF format?
5. Developed CERIF4REF plug-ins for ePrints (Southampton), DSpace (Edinburgh) & Fedora (King’s)
   - Generate xml in CERIF4REF format to allow import into other repositories or CRISs
6. Demonstrator IRs at King’s and Southampton
   - Can we assemble, supply and exchange such data?
CERIF and CERIF4REF schemas

Common European Research Information Format (CERIF)

• dev. in 1980s & maintained by euroCRIS, a not-for-profit organisation
• Intended to enable data exchange between research information systems
• Available in SQL and XML versions

CERIF4REF application profile

• Simplified version of CERIF tailored for REF
  • Less elements than full CERIF scheme
  • Cardinality changed to reflect REF requirements
• Guidance on information to provide in each field (e.g. controlled vocabulary, suggested terms)
• XML scheme
  • Single XML schema containing all elements
Mapping data to CERIF4REF

• Worked with 5 institutions to trial CERIF4REF
  • Goldsmiths, Kingston, Reading, Leicester, University of the Arts, Ulster
• Asked to map their currently-held data to the CERIF4REF data dictionary
• Data obtained from:
  • Staff HR, finance, student information, research database, publication system (usually eprints), pre-award db, grant db, and 3rd parties (usually web of science)
Mapping data to CERIF4REF: Experience & Challenges (1)

General concerns:
• Concerns regarding the use of EPrints to store sensitive data

Quality of data:
• Some data elements require cleansing or re-keying prior to export to C4R
  • Citation style for author/contributor in repository was inconsistent with requirements
• Some information held by institution was incomplete, e.g. sponsor data & funding source in student record db
• Manual intervention was required to map some repository values to different C4r fields depending on output type
• May not be one-to-one mapping of repository to REF output types
Mapping data to CERIF4REF: Experience & Challenges (2)

**Interoperability:**

- Very little interoperability exists between corporate systems
- Inconsistency of metadata labelling between systems
- Different IDs used for staff in some systems
- Not all data in a correct or transferable format, so must be restructured
Further changes to EPrints

University of Southampton has performed additional work to enhance data model:

**Pre-CERIF EPrints**
- Traditionally, EPrints has possessed a “flat” structure - EPrints object – Metadata & 1+ optional files.
- Project and funding information held as MD elements in schema

**’CERIFied’ EPrints**
- Repository Objects – Project and Funding Organisation are objects in own right
- Able to create relationships between an EPrint object and its parent Project / Funder
Conclusions

General recommendations:
Consider the end goal:

- **Who**: Who currently uses your metadata? What audience is being to emerge?
- **What**: What functions will the metadata achieve? How will it help your audience?
- **How**: What can you achieve using resources available? What are the implications for how you capture metadata?

**CERIF4REF and the Creative Arts:**
- Determine the research outputs that will be included in your REF submission
- Review the REF requirements and determine information sources from which metadata may be obtained
- If planning to use digital repository as a basis for your REF submission, visit the R4R web site to find the CERIF4REF data model & schema, EPrints/DSpace/Fedora plug-ins, and case studies

- [http://r4r.cerch.kcl.ac.uk/](http://r4r.cerch.kcl.ac.uk/)
Thank You for your attention

QUESTIONS?

Gareth Knight
gareth.knight@kcl.ac.uk