

# Metadata topics

DELOS brainstorming, Juan les Pins 2005-12-05

*Traugott Koch, UKOLN*

I Introduction

II Metadata development needs

III Specific metadata requirements for digitized documents and preservation

---

## I Introduction

- List of important research and development requirements focused upon digitisation of documents/objects and their subsequent use
- Metadata is orthogonal to the other aspects reported today: needed to make the other efforts possible
- Metadata approaches are highly heterogeneous for different purposes like digitisation, stewardship, preservation, multimedia or multilingual objects, eScience or eLearning. Metadata for interactions between entities, for usage data, user data, choices and behaviours Metadata for services, collections, institutions, people etc., for other than solid information objects  
Widely different types of metadata: descriptive (discovery); administrative (incl. rights, access, technical, management); and structural (incl. context, presentation)
- Research and development needs are tightly linked and hard to prioritise, since we are preparing a comprehensive practical task, a mass digitisation effort and global access to the documents
- A recent evaluation of a German national digitisation programme during the last couple of years reveals clearly insufficient metadata practice, endangering the usage of the digital documents, not to speak of their preservation:

33% of the objects had no metadata at all, 33% bibliographic metadata only, 10% had both bibliographic and subject metadata (rest: no information)

Less than a third of the metadata was digital

Recommendations: a) minimum level of metadata required, based on OAI b) develop clear guidance reg. level of metadata and subject indexing

## II Metadata development needs

- **Data model, schemas**

- Harmonisation of data models for interoperability between diff. metadata-schemas, as started between the Dublin Core Abstract Model and IEEE LOM/IMS
- Generic profiles for discipline/purpose (e.g. eScience, preservation)
- New schemas needed for increasingly interactive contents
- Metadata methodologies allowing packaging and repurposing, derivative and aggregated works, recombinable content
- Application Profiles for consistency and interoperability
- Common principles for XML-binding and other encodings
- Links to OAIS and similar frameworks; develop common reference models for metadata with the same purpose and application area
- Schema creation tool
- Metadata schema registries and (semi-automatic) schema transformation (crosswalks)  
Automated import and export of schemas
- Multiple cludgy crosswalks tend to break the semantics
- Format registries
- Terminology registries
- Role of syntax and vocabulary encoding schemes and terminologies/authorities

- **Creation**

- Handle multiple metadata creation and repository environments; multiple metadata formats
- Metadata for entire life cycle of objects

- Embedded in routine workflows (repository ingestion) and organisational patterns
- Embedded/improved tools (into authoring, content managing, learning object managing tools, scientific machinery etc.)
- Support diff. metadata profiles for diff. collections
- Cooperation with content generation software vendors
- Differentiate provenance of metadata, trust, validation
- Judge degree of format adherence
- Automatic/semi-automatic creation/extraction: more research needed
- Research in metadata propagation and inheritance between related objects
- Enhance object metadata with author profiles and vice versa
- Exploit existing sources of metadata fully; exploit implicit metadata; extract from document, context, related content, author information, usage context and feedback
- Automatic metadata generation for non-textual resources
- End-user creation (social tagging)
- Role of manually created metadata; integrate human and automatic processes
- Granularity problem, e.g. for datasets
- Identify genres/document types, formats
- Include content standards (vocabularies, name authorities etc.) into metadata generation applications: tools supporting indexing, classification, KOS, ontologies
- Support for multiple controlled vocabularies
- Experiments with semi-structured metadata
- Automatic quality control
- Versioning

- Bibliographic relationship control: Work-expression-manifestation

## ● **Maintenance**

- Assure metadata update
- Tool for detecting and reporting changes to resources
- Persistent identifiers (linking, citation), different standards, identity of objects
- Digital curation of metadata itself

## ● **Quality enhancement and enrichment**

- Esp. of automatically generated and harvested metadata and for aggregator services
- Systematically evaluate experiences from large digitisation efforts: missing and erroneous information
- Sharing metadata (but: trust and policy patterns)
- Link metadata from different sources
- Aggregation, fusion vs. cross search
- Managing identity and difference (duplicates, merged, derivative)
- Add authorities (names, subjects), add vocabularies for keywords (for consistency)
- Relate names and identities across multiple data streams
- Subject mapping, interoperability
- Tools and methods needed for metadata enhancement

## ● **Packaging, complex objects**

- Need to manage complex digital assets/objects
- Preserve context
- Manage huge repositories of objects and metadata

- METS, MPEG-DIDL approaches; IMS Content Packaging Spec.; SCORM. Some are not useful for frequently changing (meta)data
- More programmatic approaches of working with multiple simple and complex objects
- Packaging issues need more investigation from a preservation perspective

- **Use, reuse**

- Seamless connection between different types of collections: books/journals; special collections and archives; research and learning materials; freely accessible web resources
- Develop rights metadata (authorization of users in open web), policies, licenses; make it machine readable
- Terminology support in other languages than English
- Merger between metadata and KO services
- Knowledge extraction from metadata
- Link between document metadata and citation indexing (ISI Web Citation Index)
- Analyse actual use of metadata by diff. types of users and use for improvements
- Social use, annotation, (collaborative web) tagging, rating, rankings, reviews/recommendations
- Investigate how formal registries and informal social tagging might eventually overlap or converge
- Address problems when reusing metadata created for multiple diff. purposes and contexts
- Use structure of the data in interfaces
- Use metadata programmatically: to 'FRBRize', to do collection analysis, to generate interesting displays
- Use metadata to create adaptive user interfaces for diff. groups of users
- Expose metadata through a variety of interfaces and protocols for searching, harvesting and search engine reuse

- Technical issues of exposing large quantities of metadata
  - Making services available from (metadata) records rather than vice versa
  - Leverage metadata functionality outside the systems they reside in today and make them useful inside new applications (web services approach), in multiple repositories, creation environments and discovery mechanisms. How to tie together (service orchestration); infrastructure for such services
  - Metadata services as web services m2m, e.g.: generation, augmentation, transformation, equivalence, crosswalking schemas and vocabularies, archiving/persistence, annotation, metadata improvement and rating services
- 

### **III Specific metadata requirements for digitized documents and preservation**

- Needs to allow re-creation and interpretation of the structure and content of digital data over time: discovery, technical rendering of objects, recording of contexts and provenance, documentation of repository actions and policies
- Scope and depth of information needed to support digital preservation: processes not known yet
- PREMIS Data Dictionary format 2005, based upon the OAIS reference model; about objects at different levels of aggregation, events, agents and rights; technical elements are still missing
- Better understanding needed of the role of metadata in supporting preservation and data curation
- Relationship between metadata for actual discovery and access and metadata for long term preservation
- Different kinds of metadata will be needed to support different digital preservation strategies
- Automatic capturing of metadata about the (complex) objects, the actions undertaken on them and about people, organisations or software controlling these actions
- Automatic capturing of provenance metadata when ingesting into repositories
- Automated preservation metadata tools needed

- Explore implications of exchanging metadata through heterogeneous digital archiving systems used for collaborative metadata management
- Quality: in the context of preservation, inconsistent, incomplete and misleading metadata must be avoided, because it will persist for a long time
- Research data and multimedia products have had specific problems with insufficient incentives for their creators to do metadata
- Hidden subjectivity and cultural bias is potentially more damaging for metadata to be used under long time. Data and its organisation must be historicised with as rich semantics and representation information as possible
- Preservation metadata must be preserved itself, migrated and described and upgraded to evolving new metadata standards. This might be more difficult when it is packaged together with the objects
- In Cultural Heritage sector: object description turns into valuable work itself
- Specific importance of contextual metadata to archivists and records managers (draft ISO standard Records management metadata): people, policies, processes and systems and the records themselves
- Provenance; legal, administrative, procedural, documentary and technological context; use history, integrity and authenticity of materials more important
- Provide for long-term access and management
- Use of metadata for collection analysis and to inform strategic decisions reg. multi-institution mass digitisation programmes: coverage, language, copyright, bibliographic units, convergence (adding new, other collections)

---

[Traugott Koch](#) (email: [T.Koch@ukoln.ac.uk](mailto:T.Koch@ukoln.ac.uk))

Created: 2005-11-30

Last updated: 2005-12-04

URL: <http://>