Universität der Bundeswehr München

Max Planck Society eScience Seminar 2008
Aspects of Long-Term Archiving

Standards and Standardization in the Context of eScience and dLTP

June 19, 2008

Peter Rödig

Institute for Software Technology
Universität der Bundeswehr München
Prof. Dr. Uwe M. Borghoff
www.unibw.de/uwe.borghoff

Universität der Bundeswehr München

Funded by nestor

nestor

www.langzeitarchivierung.de

www.longtermpreservation.de

In cooperation with Bayerische Staatsbibliothek München

BSB Bayerische StaatsBibliothek

www.bsb-muenchen.de

Outline

Universität der Bundeswehr München

- Basics of standards and standardization
- OAIS-RM as a means for organizing standards
- Characteristics of eScience and impacts to OAIS-RM elements / standards
- Assessment of current situation
- Some proposals

Do We Need Standards?

Standards can

- provide the precondition for cooperating technical infrastructure
- make concepts, products, systems, services comparable
- give consumers the opportunity to choose different solutions
- reduce the need to update (large) systems synchronously
- provide economical benefits
- increase the trustworthiness of dLTP

Why?

- Standardization processes are transparent and open
- Standardization is voluntary
- Results are documented, publicly available, and based on consensus
- Standardization Organizations (SOs) provide means to assess conformance
- Standards are based on consolidated knowledge

Drawbacks

A clear taxonomy for standards is missing, definitions depend on several factors like subject, countries, bodies/organizations

An Approach

- **Normative Stds**: prescriptive (conformance)
- vs. **Informative Stds:** provide guidance and helpful information
- **Formal Stds:** normative, require formal approval processes of a formal national or an international standards body (DIN, CEN, ISO), also called de jure Stds
- **vs. Informal Stds**: normative (technical specs) or informative (recommendations, reports), developed by a standards body or a SDO
- **De-facto Stds, Quasi Stds** (Industriestandards): dominate the market / broadly accepted (e.g. HP‘s PCL) also called Non-Stds
- **Private Stds:** normative or informative, closed membership

Drawbacks

- Many Definitions for „Open Standards“

  Openness is preferred in the context of dLTP, but some conditions are considered as too restrictive (e.g. “reasonable” in RAND)

  See the heavy debate about ISO‘s OOXML

- (too?) Many SOs / SDOs

- (too?) Many Standards and Standardization Efforts

  -> A tool for organizing standards and standardization is required

Benefit

OAIS-RM decomposes a large system and describes important concepts

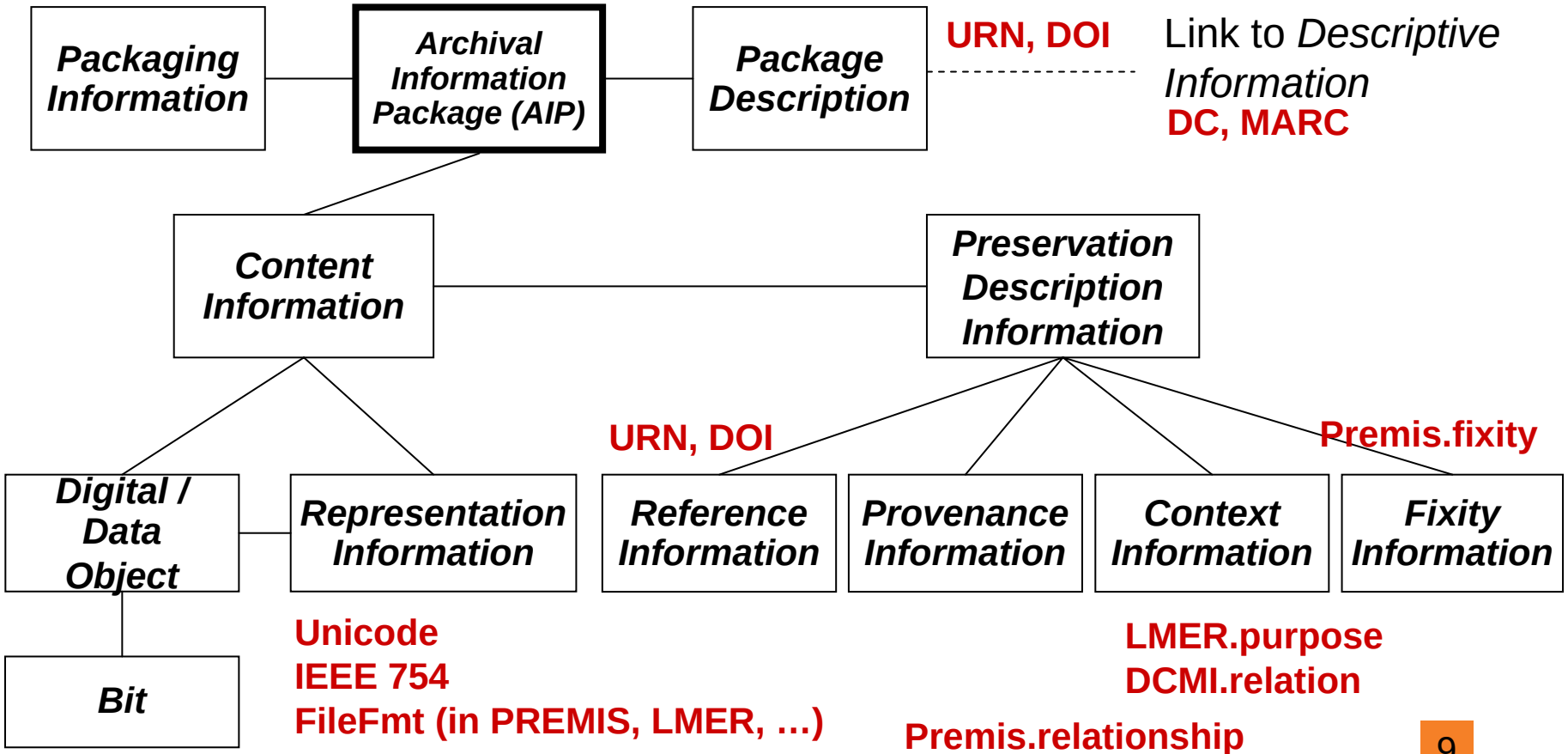
- Organizational View
  - Major tasks and roles
- Functional View
  - Major functions
- Informational View
  - Relation between data and information
  - Categorization of information

- Standards (or parts of them) can be mapped to RM-elements
- RM-elements can be used to guide (checklist) standardization

Of course, OAIS-RM does not cover all aspects

other types of IPs: Submission IP, Dissemination IP

Mets, ZIP, XFDU … more elements for organizing AIPs are available in OAIS



… in comparison with „traditional“ digital objects

- Extreme large and/or numerous digital objects, use of Grid middleware

> *Common Services, Archival Storage* (middleware, virtualization)

>> no specific dLTP-standards (as attribute in some metadata standards)

- Specific file formats (and interfaces)

> *Representation Information*

>> Premis, LMER, Pronom, GDFR

- Individual formats (designed ad hoc, without standardization, definition languages in use)

> *Representation Information*

>>Premis, LMER, Pronom, GDFR

- Specific instruments esp. for capturing signals/data

> *Provenance, Representation Information*

>>Premis, LMER, Pronom, GDFR

- Complex processing of data, e.g.

  cluster analyses, constraint satisfaction, large-scale eigenvalue problems, complex pre and post processing

> *Context*, *Provenance*, *Representation Information*

>>Premis, LMER, dLTP-Standards for registering services (UDDI ?)

- Fuzzy objects (vs. discrete digital objects)

> *Packaging Information, Reference Information, Context Information*

>>METS, Persistent Identifier

- Complex relations between objects

  (e.g. data – experiments – learning objects – scientific publications – scientific programs)

> *Packaging, Archival Information Collections*, *Reference Information, Context Information*

>>METS, cataloging (MARC)

- Scientific workflow (here, consolidating and presenting results)

> Building *SIPs, AIPs, DIPs* (*Functional* View: Processes between *Producer – Archive – Consumer*)

>>ISO 20652 (abstract co-standard for OAIS-RM for *Producer – Archive* interface)

Practices but no concrete standards yet?

… more details in a report of the nestor-WG *Grid/eScience and LTP* soon

Current dLTP-standards provide a good basis for eScience, e.g.

- frames for describing complex *Information Packages*
- architectures for maintaining (application-independent) *Representation Information* in registries

But

- Core concepts in dLTP are not defined in a unified or precise way (digital object, digital entity, version, identity, migration, manifestation, …)
- The OAIS *Information Model - Representation Information* is too abstract
- dLTP standards have redundancies (e.g. for describing the structure of *Information Packages* and *Information Collections*)
- dLTP standards are strongly oriented to computer files (refinement is on the way)
- dLTP standards have no means to describe complex scientific processing e.g. for numbers (accuracy problems) or complex computing environments (*Provenance*, derivation of *Information Packages*)

Some Statements and Suggestions

1. Standardization can be approved by reference models
2. Large systems require an architectural approach (with adequate layers of abstraction)
3. Standardization needs input (use cases) and feedback from practice

- Designing reference models and architectures for dLTP

Starting point

- OAIS-RM

Good ideas in

- OASIS‘s SOA Reference Model (explains different roles of architectures)
- OGF‘s OGSA esp. in GLUE (formalizes an architecture, can provide CV)
- Practices and private standards !?

Some Suggestions

- Defining parts of an architecture with relevance for eScience and „traditional“ dLTP
  - Concepts for identity (*Reference Information*)
    what is referenced (resource, *AIP*, ...), what will happen when migrating, …
  - Concepts for container (*Information Packages*) and their profiling
    how to navigate in complex objects, what will happen when migrating, …

  (not on the level of a specific grammar)
- Looking for OGSA-elements that
  - fit into the frames of current dLTP-standards or
  - refine OAIS-RM esp. *Common Services* (e.g. data service)
- Extending / populating existing or planned registries
- Identifying and documenting use cases

Thank You !

CV	Controlled Vocabulary
GDFR	Global Digital Format Registry
LMER	Long-term preservation Metadata for Electronic Resources
METS	Metadata Encoding and Transmission Standard
OASIS	Organization for the Advancement of Structured Information Standards
OGF	Open Grid Forum
OGSA	Open Grid Services Architecture
PCL	Printer Command Language
RAND	Reasonable And NonDiscriminatory
UDDI	Universal Description, Discovery and Integration
XFDU	XML Formatted Data Unit

ISO 20652:2006 Space data and information transfer systems - Producer-archive interface - Methodology abstract standard

ISO 14721:2003 Space data and information transfer systems - Open archival information system - Reference model