



Relevance, Standards and Usage of Metadata for Electronic Language Resources

Daan Broeder, Peter Wittenburg

MPI for Psycholinguistics

CLARIN Research Infrastructure

HT: there is not (yet) one agreed descriptive system for LRT. Let's
limit the damage!



Library History

- concept of descriptive metadata is of course very old
 - library catalogues were used to easily manage and find books stored somewhere on the shelves
 - some liked the catalogues – others liked to look at the book instances
 - these catalogues typically had very limited information
 - finding the right book (title, author, year, etc)
 - quick inspection (citation, older versions, statistics, etc)
 - managing the library holding (overview, reorganization, missing, etc)
 - not the place for deep characterization
 - in some catalogues content classification
 - genre
 - subject (LCSH, IconClass , ...)
 - libraries the first to introduce/push electronic catalogues and exchange formats (MARC, etc)
 - Dublin Core to describe any authored web-resource was pushed forward also by librarians.
-



Motivation in Language Resource Domain

- Constantly more language resources are created of all types.
- At MPI about 500.000 digital objects deposited from a large group of researchers independently of each other with a high annual increase
- The sheer quantity requires new methods to prevent

Digital Chaos or Data Cemetery

1. need good and stable **repositories/archives**
 2. need a good **Descriptive Metadata infrastructure**
- Several of us realized this
 - Early approaches
 - TEI header tags (deep descriptive intention) were used in various projects (Dutch Spoken Corpus)
 - CHILDES annotation file header tags (search, filtering etc)
 - ...
-



Initiatives for Descriptive Metadata for LRT

- Dublin Core MD initiative for all types of authored web-resources
 - 1998 TEI header
 - May 2000 ISLE MD White Paper (IMDI) presented at LREC in Athens & establishment of an IMDI working group
 - May 2000 LREC necessity of language classification system (Ethnologue) now an ISO standard
 - December 2000 Presentation of the OLAC initiative
 - 2000 DFKI/ACL Registry of tools

 - important activities in other fields
 - LOM: DMD for learning objects
 - MPEG7: complex integrated approach (DMD + content)
 - ISO 19115: geographic information
 - Indecs
 - ...

 - social tagging as alternative for expert metadata, but usability for our domain may be limited
-



Functions of Descriptive Metadata for LRT I

Differences in the approaches wrt to different interest groups

- **users**

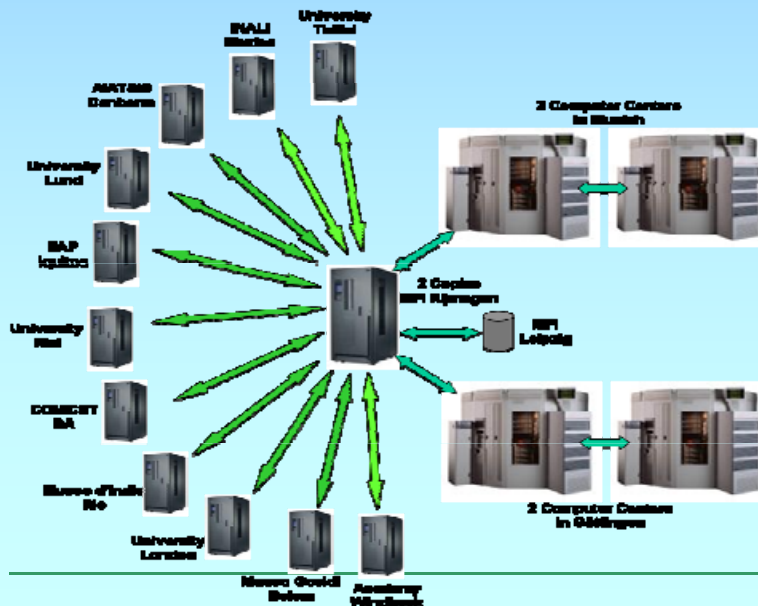
- search big catalogues with a large number of descriptions
 - browsing through linked hierarchies or networks of DMD
 - faceted browsing as a combination
 - geographic browsing based on GIS coordinates
 - quick inspection of metadata to check suitability
 - virtual collection building and workflow creation (process journal)
 - creating relations between LRs of various sorts
 - creating different views including dynamic web-sites
 - research questions vs. discovery
 - *give me frequencies of correct usage of 3. person plural inflected form for children of different age and sex*
 - *give me lexicons for Trumai*
 - granularity
 - *let me find a specific individual object*
 - *let me find a corpus*
-



Functions of Descriptive Metadata for LRT II

• depositors/managers

- canonical hierarchy according to linguistic criteria and resource bundling (container building)
- for resource management (migration, moving, etc)
- for simple access and license management
- adding valuable information/knowledge about resources
- for copying parts (access, long-term archiving)



Example Scenario:

- all copied to computer centers
- only parts exchanged between MPI and regional centers in both directions



DMD Infrastructure Components (until now)

- metadata provider <-> service provider
 - one major difference: DMD Data vs. DMD Service Provider
 - DMD Service Provider has no resource management task
 - the DMD specification
 - a schema (flat or structured - until now is one of the main pillars)
 - a vocabulary of descriptor elements with key-value pairs
 - per elements value sets (closed, open, semi-closed)
 - special profiles to include new sub-disciplines
 - the tools
 - editor, browser(s), search engine (structured vs. unstructured)
 - DBMSs. (relational or XML based)
 - OAI-PMH protocol (gateway, harvesting)
 - linker and virtual collection builder
 - view generators
 - APIs (Web services: SOAP & REST)
 - ..
-



DMD Experience

- some initiatives have done an excellent job
 - IMDI, OLAC have stabilized and offer services
 - DC moved from 15 broad categories to qualified concepts
 - vocabularies are registered (community sites, ISO DCR)
 - OAI PMH is widely accepted for metadata exchange
 - XML harvesting as a less expensive alternative is accepted
 - but total coverage is not at all sufficient
 - too few repositories are ready/willing to participate
 - DMD usage is not at all satisfying (see IMDI usage*)
 - necessity not believed despite evangelization
 - DMD generation costs money, but is not budgeted
 - some researchers still don't want to share
 - some researchers would like to participate, but ...
 - lot of legacy material – how to get that in?
 - DMD is open – some have ethical/political problems
 - user friendliness (what is this?) to be improved
 - not all functions supported
-



DMD Lessons learned

- schemas are secondary - let everyone create his/her own schema
 - primary are registered and suitable vocabularies and persistent IDs
 - a registry for schemas to allow re-usage and look-up
 - need a flexible component based framework for DMD (similar to LMF)
 - a REQUIREMENT to use registered vocabularies
 - need to support localization and sub-discipline terminology
 - a registry allowing to re-use existing schemas or blocks (schema fragments)
 - easy registration of new schemas (using registered vocabularies)
 - full support of PIDs at all relevant levels: concepts, resources and other metadata
 - possibility to register useful relations between concepts (pragmatic ontologies)
 - a next generation tools should support such a framework
 - need thorough studies of resource types / descriptor sets per type
 - need to include web services so others may interact with it.
-



Standards and other Trends

which standards/suggestions are there

- ISO TC37/SC4: ISOcat as DCR standard on the way
 - of course reuse trustful registries such as DCMI
 - W3C, ISO, IETF: PID standard
 - TEI ODD component framework
-
- which standards/suggestions are missing
 - exhaustive LRT taxonomy and description per data resource type
 - a feasible suggestion for WS description (UDDI, ebXML did not work)
 - an accepted model for new generation DMD
 - DMD is in the focus of large initiatives such as DRIVER
(European project to create a Digital Repository Infrastructure)
 - **will someone take care?**
 - **in the CLARIN project this all will be one of the main issues**
 - **a flexible component model for MD is on the CLARIN list**
 - **poster on Thursday P15**
-



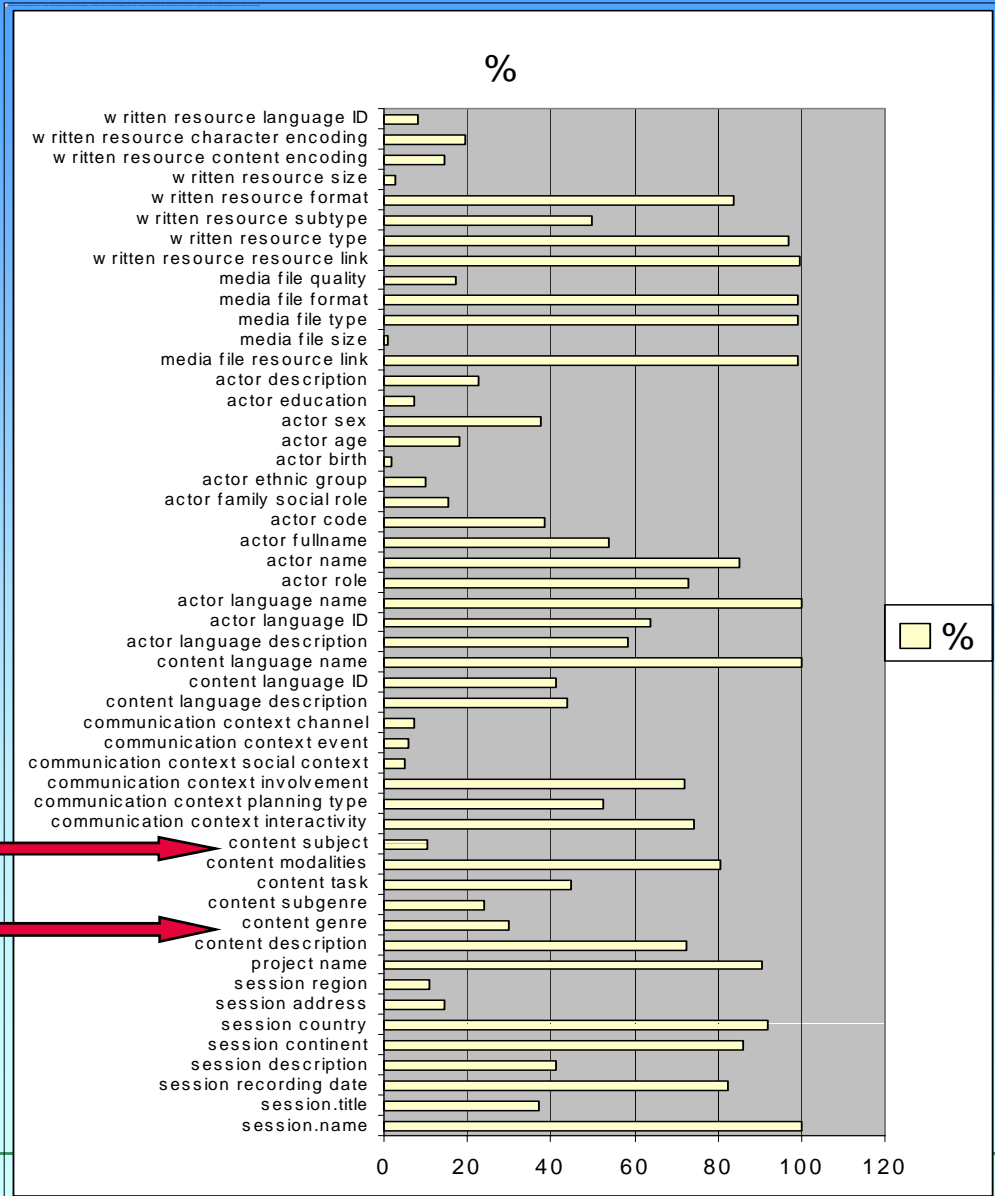
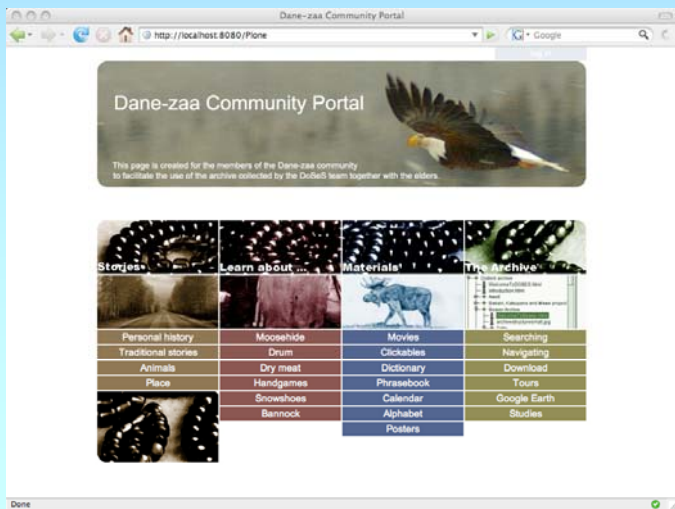
Thank you for your kind attention.



IMDI Usage

IMDI statistics on 27.000 records:

- many creators did not use content fields (Genre, Subject)
- difficulties with classification, laziness – why should I invest time, etc
- now after 6 years special web pages with dynamic REST-based content generation, motivation increases



d g2

they already know the data
time investment is for other users
data is considered own property and not that of the funder or research community

REST -> more IMDI ????
but it makes them more aware of the
possibilities of metadata
broeder; 23.05.2008