IMDI & Endangered Languages Archives

Heidi Johnson / AILLA

Acronyms & URLs

- IMDI = International Standards for Language Engineering MetaData Initiative: http://www.mpi.nl/ISLE
- MPI = Max-Planck Institute for Psycholinguistics: http://www.mpi.nl/
- DOBES = Documentation of Endangered Languages: http://www.mpi.nl/dobes/

Overview

- Goal: bottom-up design of a metadata schema for resources archived for DOBES.
- Considerations:
  - DC elements too shallow & fragmented.
  - Want to be able to "bundle" resources together.
  - Want to include all the information concerning a resource in its metadata schema.

Bundles of materials

- Multi-part resources:
  - Audio/video recording of a speech event; e.g. narration of a traditional myth;
  - Transcriptions, translations, & annotations;
  - Photographs, additional tracks, etc.;
  - Multiple formats are archived: .wav & .mp3; pdf & txt…

A problem for the DC/OLAC model:

How can we keep related resources together & make sure users get all the parts they need?

The IMDI Session Schema

- Describe a single time-bounded recording, plus derivatives (e.g. transcriptions).
- The schema is large & highly structured.
- Sub-schemas are "shareable" with other schemas, like the Written Resources Schema.
- Every sub-schema has a Description field
- Every sub-schema has customizable Key/Value pairs.

Session Schema: the big pieces

- Session info: Title, Abbr title, Date & Place.
- Project info: Title, Contact info.
- Depositor (Collector): Name & Contact info.
- Participants sub-schema
- Content sub-schema
- Resources sub-schema
- References

Participants sub-schema

- Name, nickname
- Role: = OLAC Role attribute
- Social/family role: parent, shaman…
- Age, sex, ethnicity, education level
- Place of origin
- Language(s): first given is native language.

Content sub-schema

- Modality: speech, writing, gesture
- Language(s): = Subject.language
- Genre: conversation, verbal contest, interview, meeting/gathering, riddling, consultation, greeting/leave-taking, humor, insult/praise, letter; procedure, recipe, description, instruction, commentary, essay, report/news; narrative, oratory, ceremony, poetry, song, drama, prayer, lament, joke; textbook, primer, workbook, reader, exam, guide, problem set; dictionary, word-list, grammar, sketch, field notes
- Communication context: elicited/non, planned/unplanned, etc.

Resources sub-schema

- Separate sub-schemas for different media. (AILLA conflates these.)
- All files:URL, size in bytes, format, access rules.
- Audio/video: quality, recording condition
- Text:
  - Character encoding, content encoding
  - Transcription & translation information
  - Language = DC Language.
  - Anonymous (use nicknames only)

MPI Implementation

- Hierarchical file system, XML files.
- Corpus Browser & Metadata Editor (PC)
- Elan: time-aligning annotation tool.
- Allows the researcher to create & manage a corpus in the field, & come home with ready-to-archive data.

AILLA Implementation

- Relational database.
- PHP Internet interface: metadata editor, search, display/download resources.
- Graded access system & user registration to protect resources.

IMDI - OLAC mapping

- OLAC terms are a subset: not everything has to be mapped
- Tricky part will be Genre: IMDI Genre conflates OLAC Linguistic data type & Linguistic discourse type
- Missing from IMDI: dataset, Linguistic field
- Missing from OLAC: teaching materials, literature (not strictly linguistic Types)

Summary

- IMDI schema includes all the info that documentary linguists want.
- It doesn't need to cover other subfields, e.g. speech recognition.
- IMDI protocols support bundling, a key consideration for AILLA.

Levels of description 1

OLAC			
Endangered language archives	Speech recognition data	Theor. papers	Language acquisition
AILLA DOBES Rausing?		ROA RRG	

Levels of description II

- Interoperability between AILLA ~ DOBES is desirable:
  - Common datatypes, resources
  - Overlapping pool of researchers (depositors)
- Interoperability between AILLA & every other linguistic archive on earth is unnecessary!

The moral of the story

- Subfields can & should define metadata schemas that cover their subjects the way they want.
- Search engines should operate at different levels of compatibility:
  - coarse search across different subfields (OLAC)
  - fine search across similar archives (AILLA, DOBES)