Understanding and Creating Metadata for Language-Resources

Alexander König
(adapted from J. Ringersma & P. Withers)
The Language Archive
Max Planck Institute for Psycholinguistics
1. What is metadata?
2. Why metadata?
3. Some metadata schemata
4. Arbil metadata editor
What is metadata?
What is metadata?

Metadata is “transcendental”

- Data about data
  (It is the ‘who, what, where and when’ of a document)
- Structured data about data
- *Internet*: machine readable data about data

Metadata is data describing a (set of) digital resource(s)
What is metadata?

Structured
in a standardized fashion using a metadata model

Metadata model
Dublin Core, OLAC, IMDI, CMDI
• Metadata scheme (elements, structure)
  • Metadata controlled vocabulary
Why metadata?

Why should you create metadata?
Why metadata?

You do not want your data to float in the cyberspace and get lost!
Why metadata?

Creation of data bases according to metadata structure

(Re) Finding resources

Using free text “key words” (Google like search)

Those who don’t take care over their metadata are doorless.

Content is expensive to create.

Just what good is your content if the people who need to read it can’t find it?
Language resources that make up corpora:

- (Digital) video or audio recordings, photographs
- Digitisations of images used as stimuli
- Transcription files
- One or more analysis files
- Field notes and experiment descriptions
- Lexica
### Dublin Core (DC) Metadata Set

<table>
<thead>
<tr>
<th>Content</th>
<th>Intellectual Property</th>
<th>Instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Creator</td>
<td>Date</td>
</tr>
<tr>
<td>Subject</td>
<td>Publisher</td>
<td>Type</td>
</tr>
<tr>
<td>Description</td>
<td>Contributor</td>
<td>Format</td>
</tr>
<tr>
<td>Language</td>
<td>Rights</td>
<td>Identifier</td>
</tr>
<tr>
<td>Relation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Metadata Schema - DC

DC example:

Content: DC.Title = “The white tiger”
         DC.Language = “English”

IP: DC.Creator = “Aravind Adiga”

Instance DC.Format = “print”
                   DC.Date = “2008-04-22”
Metadata Schema - DC

When to use DC:

**Interoperability:**

You need to offer your data to other communities using commonly understood semantics

You only need a core description

... and find the DC vocabulary/names acceptable

You only have limited resources and can only manage to enter a few fields.
Special to Language Resources

- In the linguistic domain often *clustered* resources
- Clustered because they refer to or result from the *same linguistic event/performance*.
- In IMDI terminology: *session* or *resource bundle*
Session or ‘Resource bundle’ concept:

Bundle of tightly related resources
Basic unit of linguistic analysis
Described with the same set of metadata (S)

- Eye-track
- Audio
- Video
- Photo
- Annotations
- Field notes
- Documents
**IMDI Example**

**ISLE Metadata Initiative**

**Session**

- **Name**: kanea_maa-TM
- **Title**: Preparation of 'maa' "fermented breadfruit" ('Ua Pou, Hakatao')
- **Date**: 2003-09-04

**Description**

In this session it is shown how fermented breadfruit is made which serves as a basis of many traditional Marquesan dishes (e.g. 'popoi'). It shows old and newer techniques of traditional maa-preparation. In former times, 'maa' 'fermented breadfruit' was fermented and stored in large earth pits. In this documentation the fermentation and storage of breadfruit was undertaken in a recipient made out of plaited coconut leaves and banana leaves. The basic process of making fermented breadfruit remained to be the traditional way (process of ripening and peeling breadfruits etc.). In this session it was also documented how to plait coconut leaves generally explaining three different techniques used for different purposes. The documentary also depicts traditional tool making used.

**Location**

**Project** Marquesan-DOBES

**Keys**

- conversion.IMDI.1.9to3.0.warning

Unknown mapping of Genre: consultation|procedure|unspecified --> ???
- content food preparation
- food fermented breadfruit
- technique traditional

**North Marquesan** 'Ua Pou dialect

**Content**

**Actors**
IMDI Example (Location, Project)

In this session it is shown how fermented breadfruit is made which serves as a basis of many traditional Marquesan dishes (e.g. ‘popoi’). It shows old and newer techniques of traditional maa-preparation. In former times, ‘maa’ ‘fermented breadfruit’ was fermented and stored in large earth pits. In this documentation the fermentation and storage of breadfruit was undertaken in a recipient made out of plaited coconut leaves and banana leaves. The basic process of making fermented breadfruit remained to be the traditional way (process of ripening and peeling breadfruits etc.). In this session it was also documented how to plait coconut leaves generally explaining three different techniques used for different purposes. The documentary also depicts traditional tool making procedures.

Location

- Continent: Oceania
- Country: French Polynesia
- Region: North Marquesas
- Region: 'Ua Pou
- Region: Marquesas

Project: Marquesan-DOBES

- Name: Marquesan-DOBES
- Title: The documentation of the Marquesan languages and culture in French Polynesia

Contact: Gaby Cabiliz, George Teikiehuupoko (Marquesas), Edgar Tetahiotupa (Tahiti)

Description:
The project documents several different aspects of the Marquesan culture (legends, narratives, food preparation, plant medicine, fishing techniques, Marquesan trick languages, songs, dances etc.)

Keys:
conversion.IMDI.1.9to3.0.warning
Unknown meaning of Genre: consultation/inprocedure/unspecified --> ??
IMDI Example (Languages)

Languages

**Language** Marquesan, North

*Id* ISO639-3:mrq

**Name** Marquesan, North

**Description**

North Marquesan is spoken on the north-western part of the Marquesan Polynesia: MRQ is an Oceanic language of the Austronesian Eastern Oceanic branch MRQ belongs to the Proto-Central-Eastern subgroup of Proto-Eastern Polynesia (Pawley 1966; Green 1966). MRQ is most closely related to Marquesan (CMS), Hawaiian and Green

**Language** French

**Language** Tahitian

**Keys**

IMDI__1_9.Interactional consultation
IMDI__1_9.Discursive procedure
IMDI__1_9.Interactional unspecified

content food preparation
food fermented breadfruit

techniques traditional
the curriculum lesson in Marquesan takes place once or twice a week for maximally one hour.

**Actor Tei**

- **Role** consultant
- **Name** Tei
- **FullName** Tei
- **Code** Tei
- **FamilySocialRole** husband

**Languages**

**Description**

Tei knows some French and Tahitian, but rarely employs these languages.

- **Language** Marquesan, North
- **Language** French
- **Language** Tahitian

**EthnicGroup** North Marquesan

- **Age** 62
- **BirthDate** Unspecified
- **Sex** Male

**Education**

- **Anonymized** true

**Contact**

**Keys**

North Marquesan ‘Ua Pou dialect

Food preparation
IMDI Example (Resources)
What is Arbil?

ARBIL (derived from “Archive Builder”) is an application for organising research data and associated metadata into a format appropriate for archiving.
Arbil Features

There are many features in ARBIL that enable users to view and edit their data.

The data can be viewed side by side in tables and bulk edited in the same table.

ARBIL is designed so that it can be used offline in remote locations.
Main Arbil Principles

- Workflow focused
- Table views
- Drag and drop
- Bulk copy and paste
- Multiple undo and redo
Arbil Workflow

Work Flow

1a. Copy for offline use
1b. Create
3a. Edit
3b. External Editor
4. Add Resources
5. Lamus
2. Working Directories

Remote Corpus
Local Corpus
Installing Arbil

There is a link to ARBIL on the MPI website

http://www.lat-mpi.eu/tools/arbil/

Providing you already have Java installed the webstart version is the fastest way to start

Alternately there are installers for Windows, Mac and Ubuntu (Debian).