Understanding and Creating Metadata for Language-Resources

Alexander König
(adapted from J. Ringersma & P. Withers)
The Language Archive
Max Planck Institute for Psycholinguistics

Content

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?

The Language Archive

Structured

in a standardized fashion using a metadata model

Metadata model

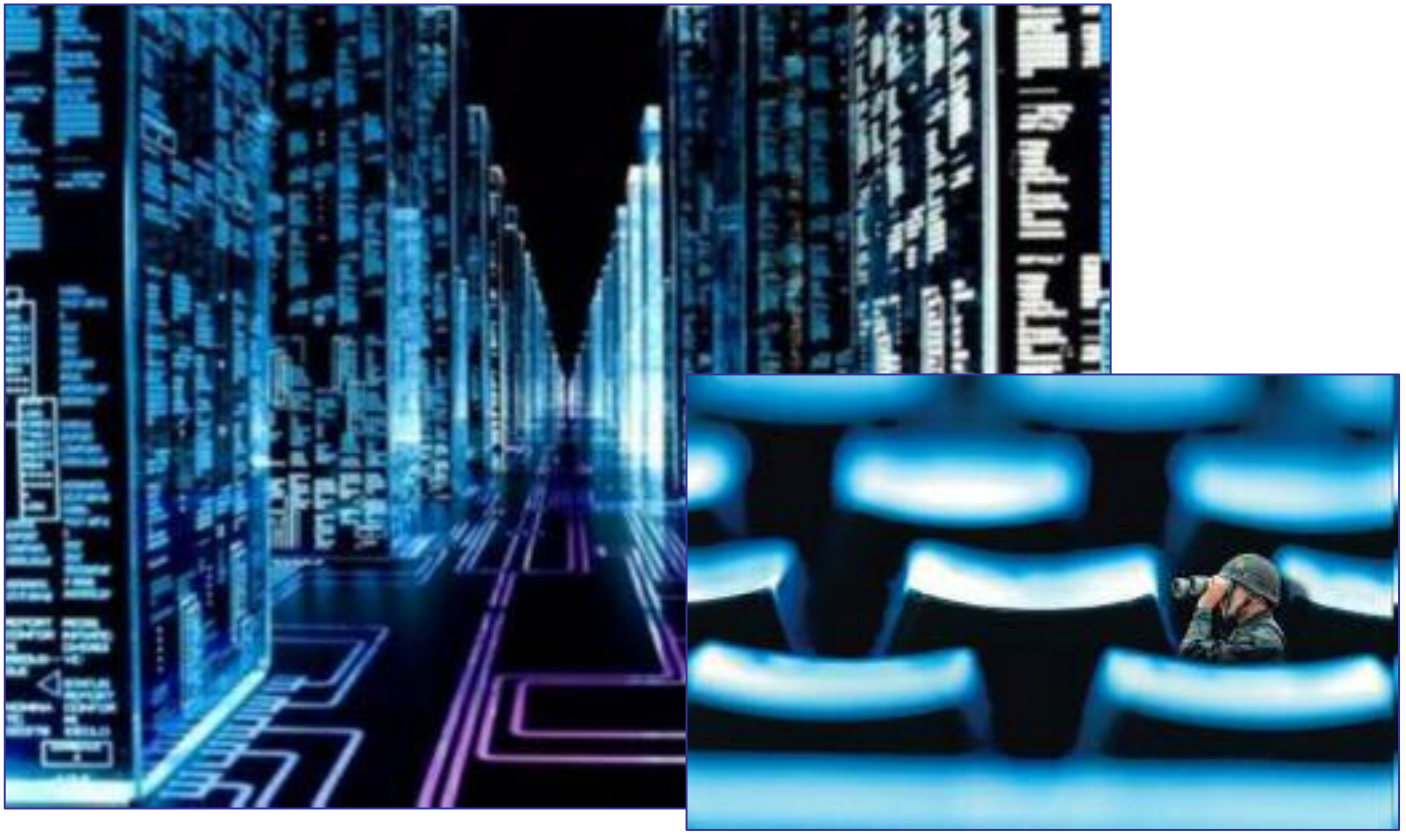
Dublin Core, OLAC, IMDI, CMDI

- 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)

Special search engines that ex… structure of metadata

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?

Metadata for Language Resources

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

Metadata Schema - DC

Dublin Core (DC) Metadata Set

Content	**Intellectual Property**	**Instance**
Title	Creator	Date
Subject	Publisher	Type
Description	Contributor	Format
Language	Rights	Identifier
Relation		
Coverage		
Source		

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.

Metadata Schema - IMDI

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**

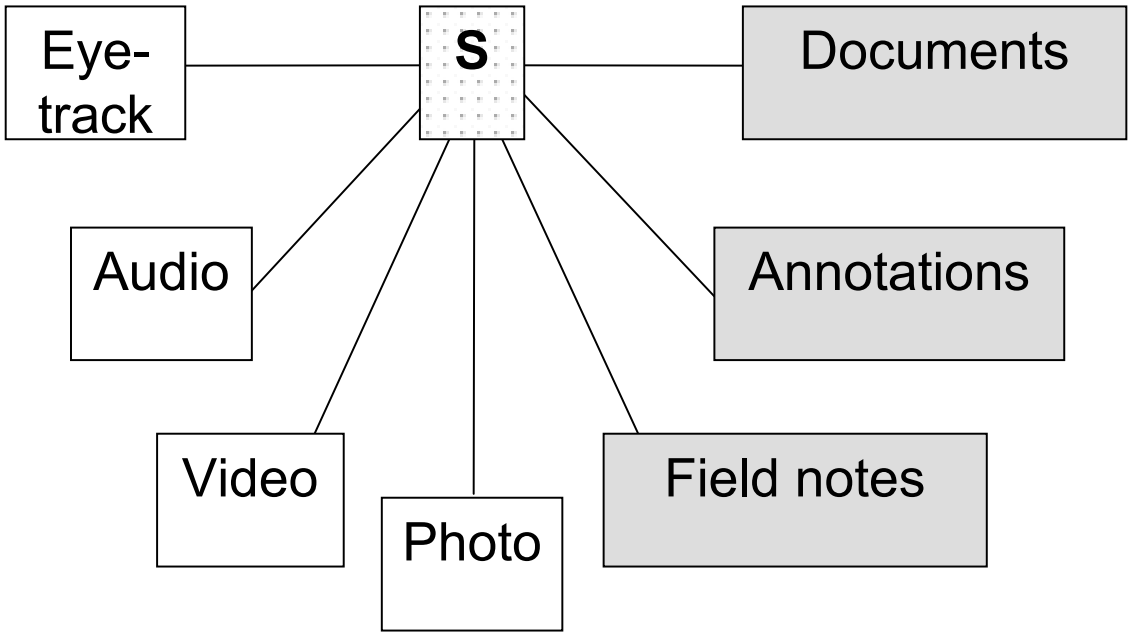
Metadata Schema - IMDI

Session or ‘Resource bundle’ concept:

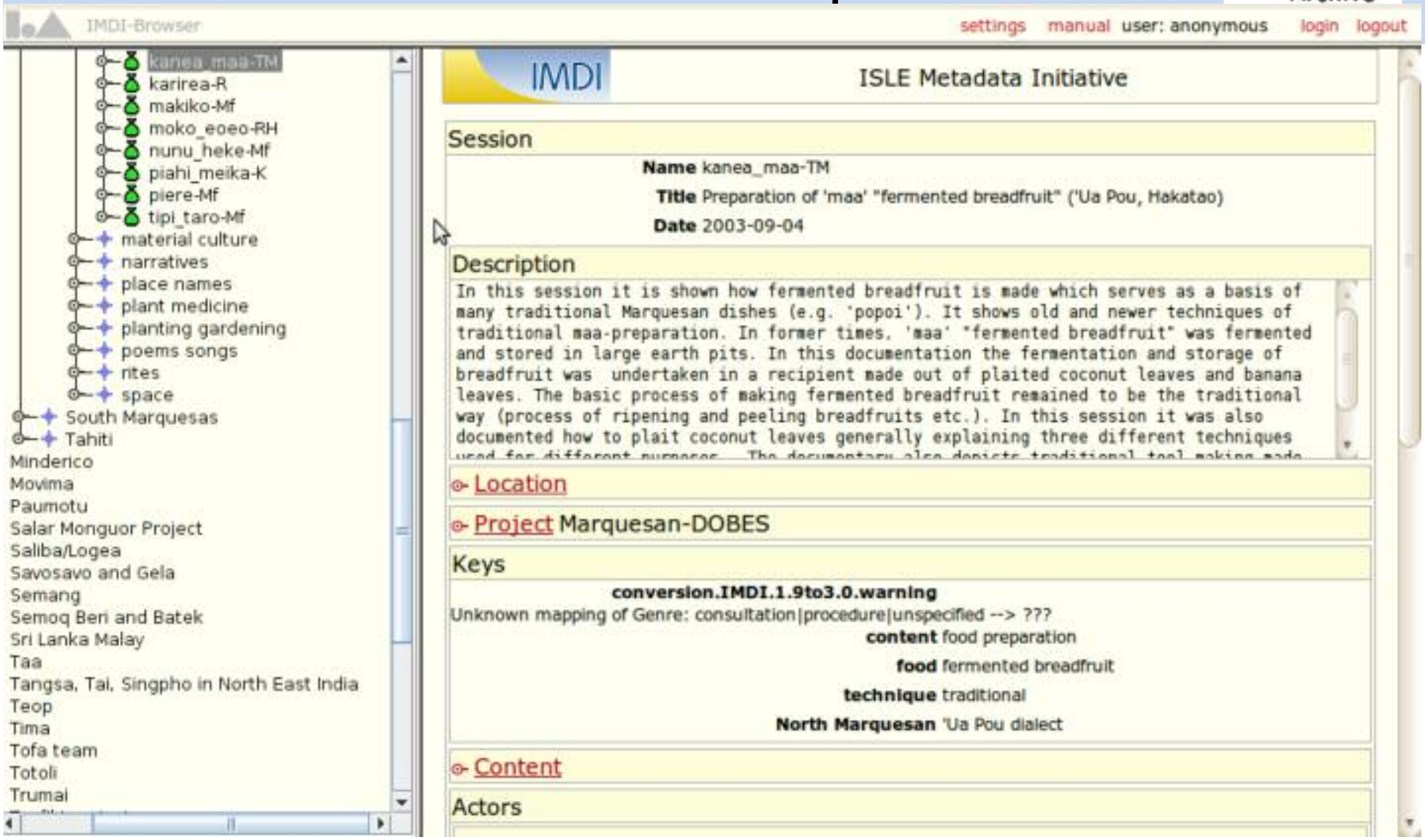
Bundle of tightly related resources

Basic unit of linguistic analysis

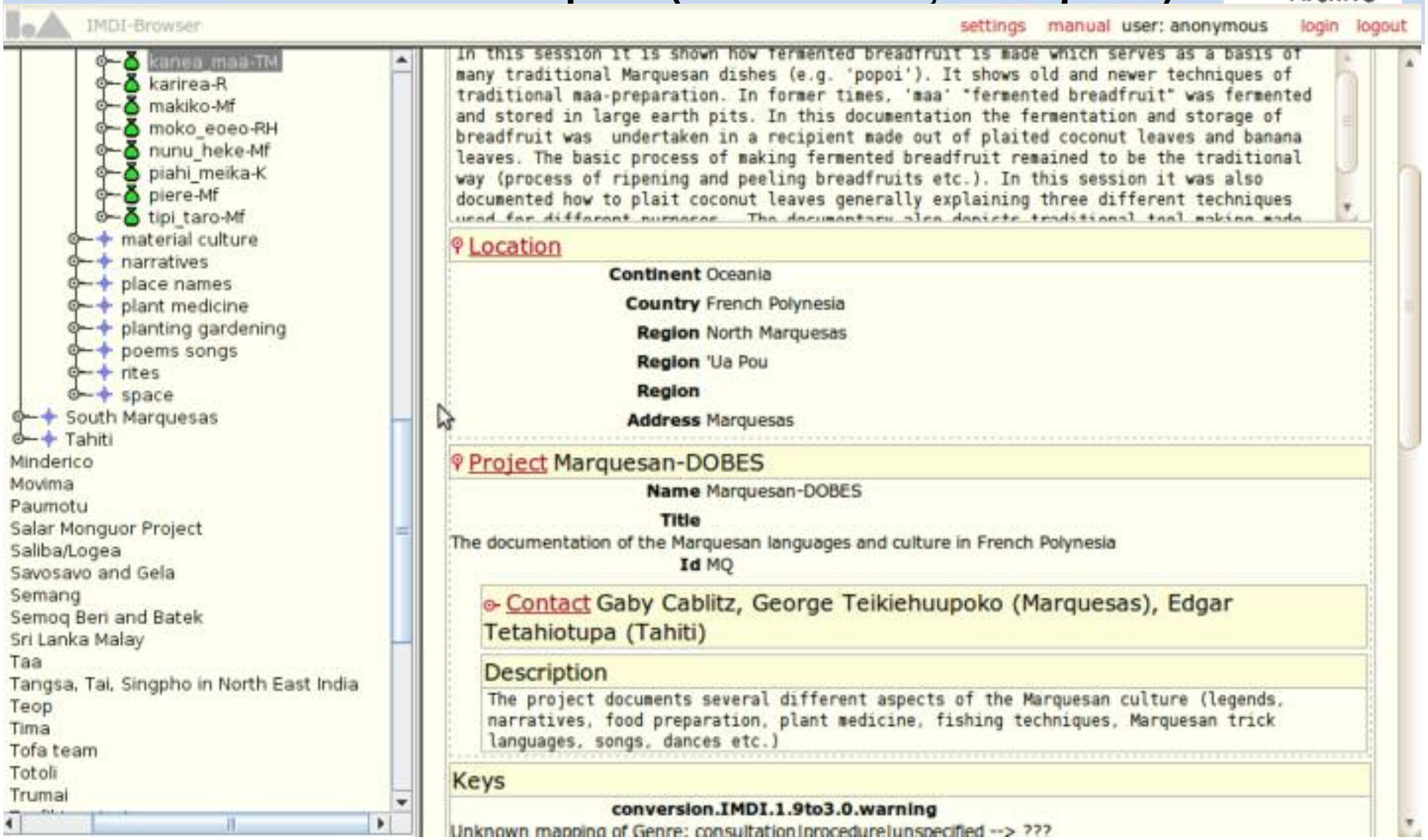
Described with the same set of metadata (S)



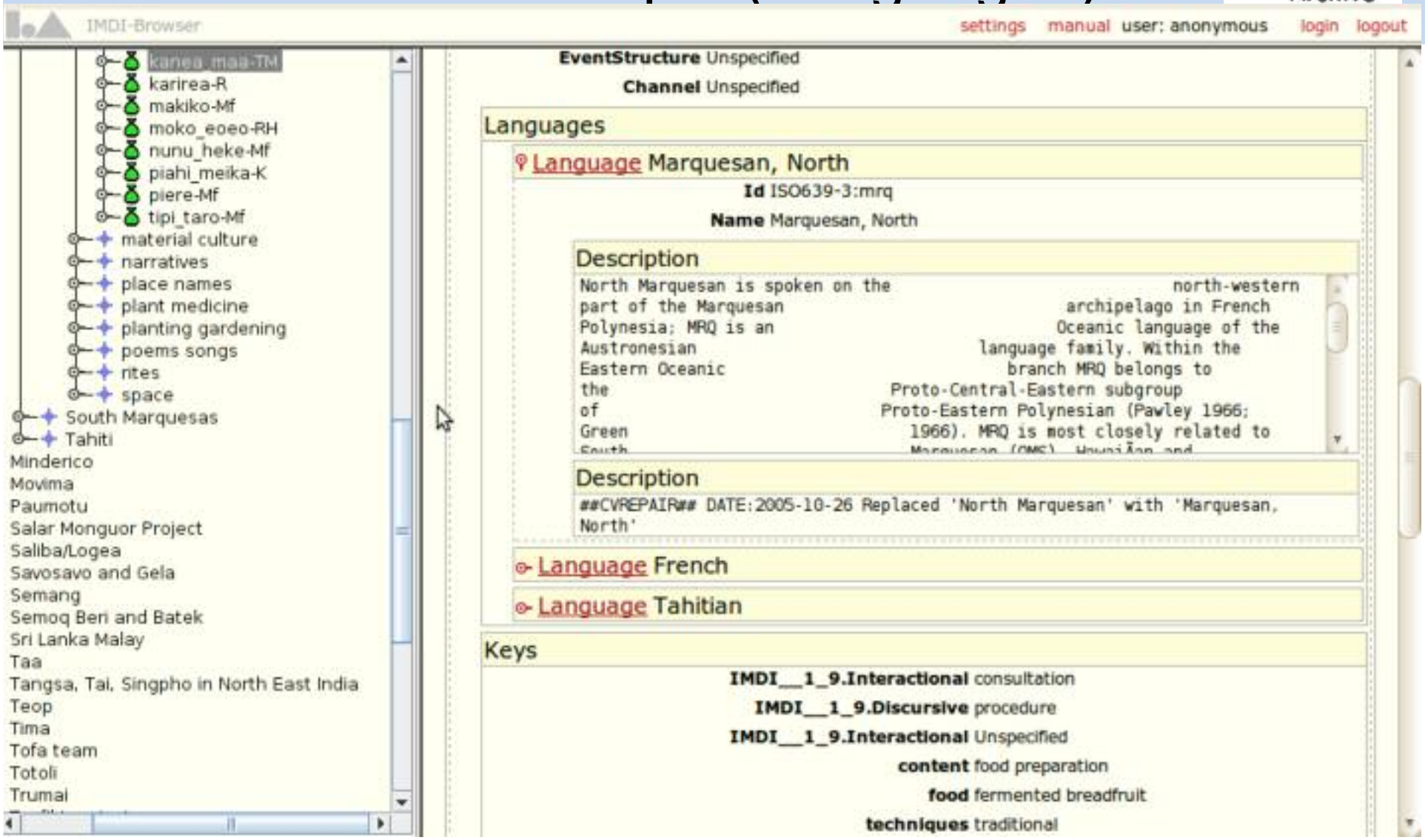
IMDI Example



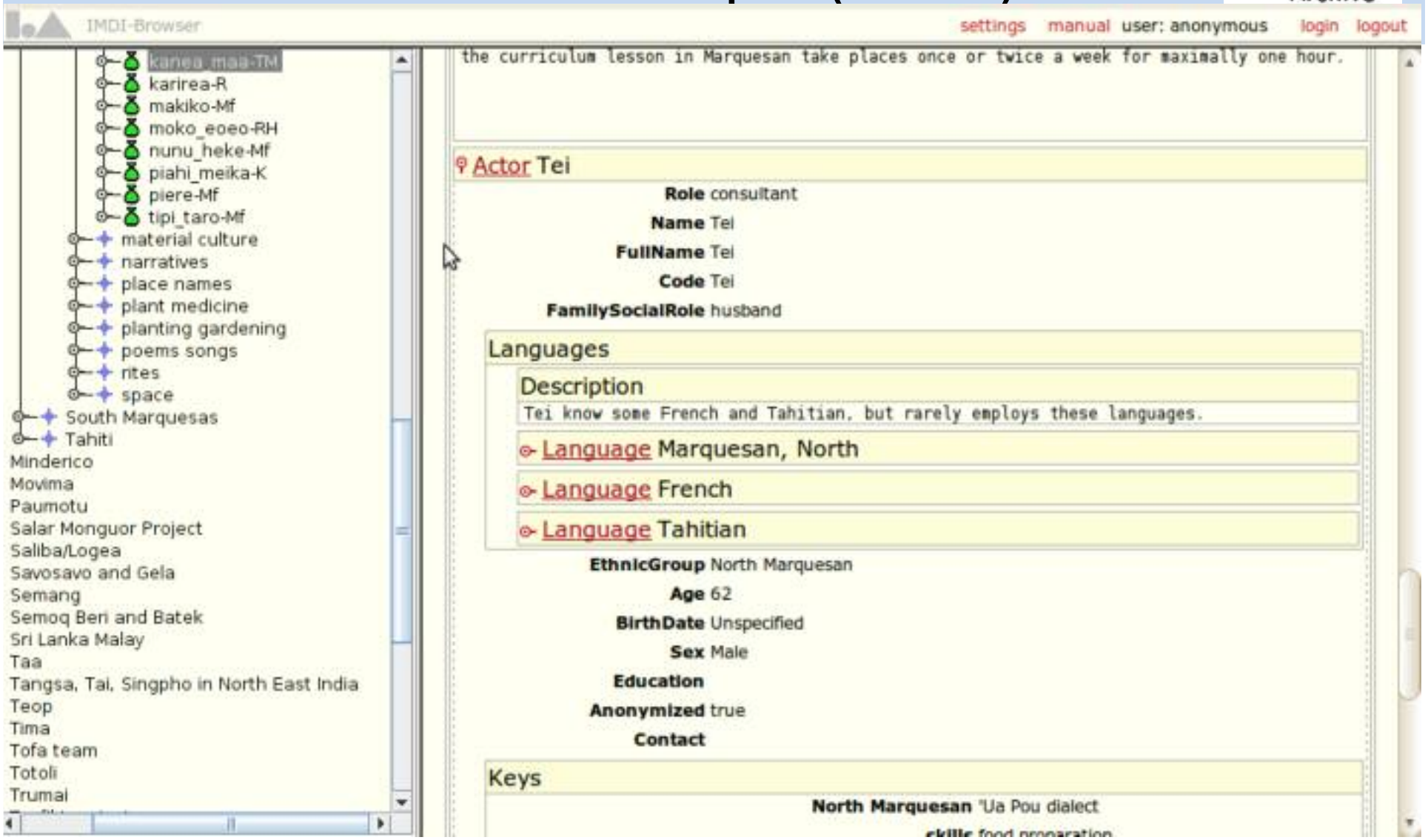
IMDI Example (Location, Project)



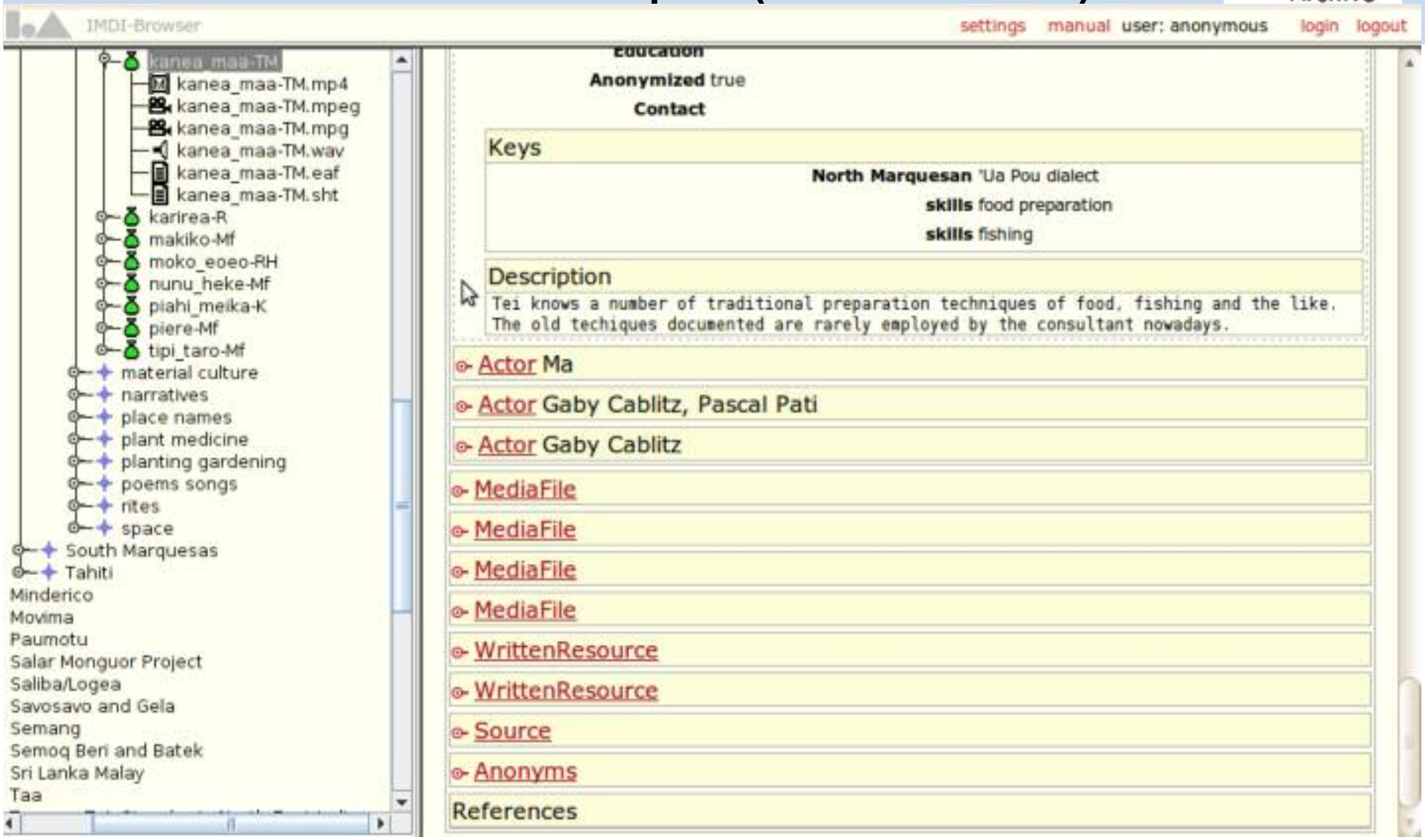
IMDI Example (Languages)



IMDI Example (Actor)



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.

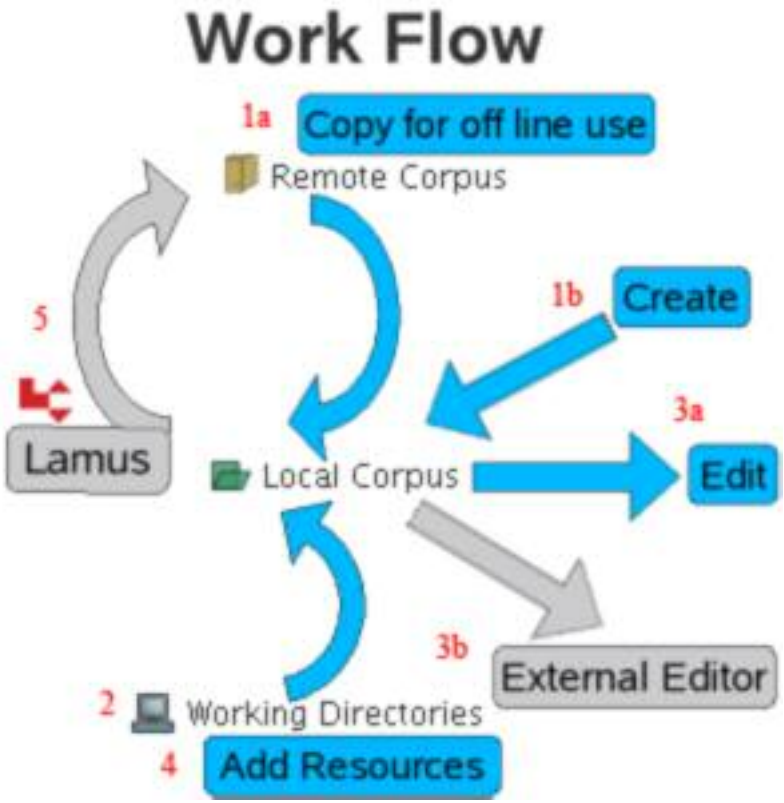
The data can be viewed side by side in tables and bulk edited in the same table.

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



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).