



MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS



The
Language
Archive



MAX-PLANCK-
GESELLSCHAFT



New Developments in Arbil Metadata Manager

Peter Withers

The TLA Language Archive,

Max Planck Institute for Psycholinguistics Wundtlaan
1, 6525 XD, Nijmegen, The Netherlands

peter.withers@mpi.nl



Overview

- What Arbil Does
- The Target Workflow
- Metadata Formats
- The User Interface
- Workflow
- Metadata Display
- Table Views
- Controlled Vocabularies
- Using Favourites
- Creating Metadata for a Resource
- Searching the Metadata
- Find / Replace and Highlighting Cells
- Installing Arbil
- Arbil in the Future
 - Marbil
 - YAAS prototype

What Arbil Does

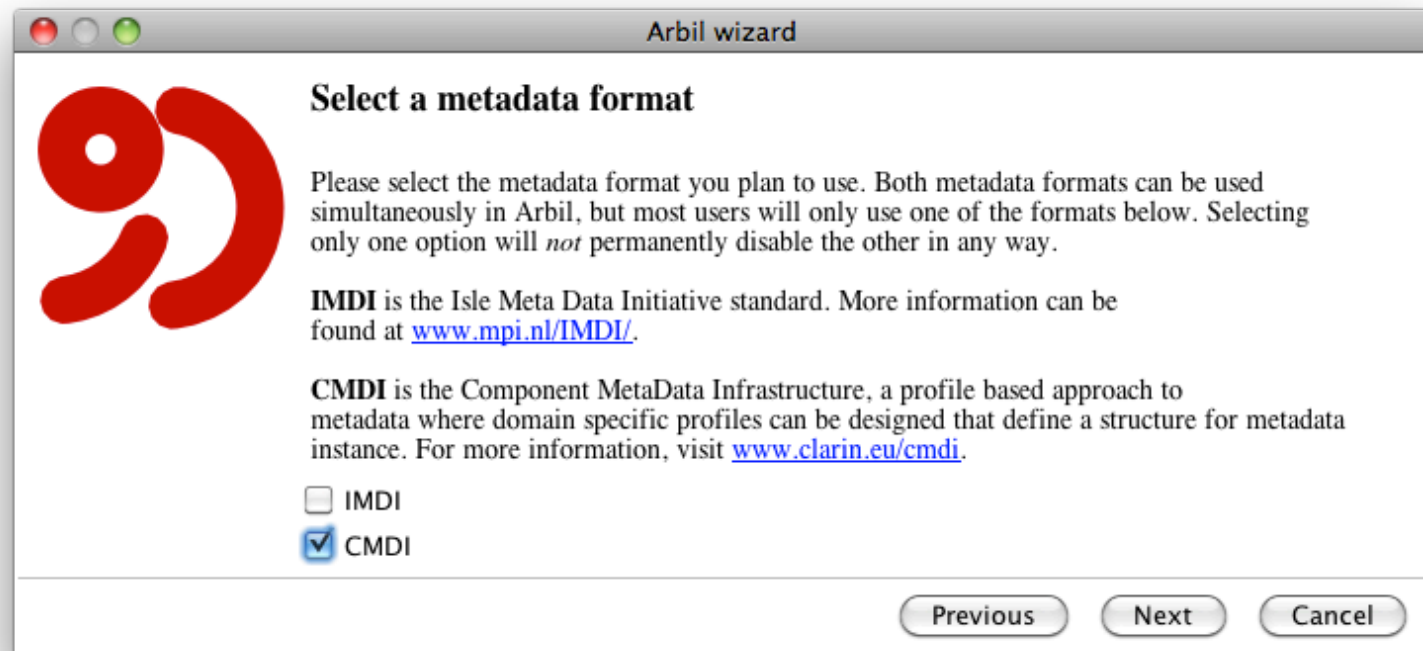
- ARBIL is an application for organising your research data with metadata such that it can be archived as opposed to buried.
- Metadata structures can be created, viewed and edited.
- Multiple metadata files can be edited at the same time.
- The local metadata tree can be searched with multiple parameters.
- The data files can be launched in the associated applications such as ELAN or Media Players.
- Many metadata files can be bulk edited in a single table.
- With the exception of accessing the remote archive, all features are available offline such as in remote field sites.
- You can enter the data as it becomes available, there is no mandatory order.

The Target Workflow

- Arbil was designed for the DoBeS community
- There was a specific workflow in use
- The users were trained in that workflow
- Curation and offline use was a core need
- Outside of the DoBeS community there are many possible workflows and different needs
- To achieve this Arbil will need to be more flexible yet still guide the user

Metadata Formats

- Both IMDI and Clarin metadata formats are supported.
- Other XML formats can potentially be supported by the use of custom templates or schema files.



Metadata Formats

IMDI (ISLE Meta Data Initiative) is a metadata format for linguistic data

- Has been around for about a decade
- Targeted towards multimodal / multimedia
- Has a single schema file with fixed metadata fields and optional key fields
- Is used for instance throughout the Language Archive in Nijmegen and associated archives

Clarin metadata is a flexible format with ISOCAT

- It is a more recent metadata format
- Each profile has its own schema files
- Developed as a part of the Clarin EU project
- Can be adapted to suit specific project needs
- The metadata fields and layout are customisable



The User Interface

Remote Corpus
View and import
metadata from remote
servers

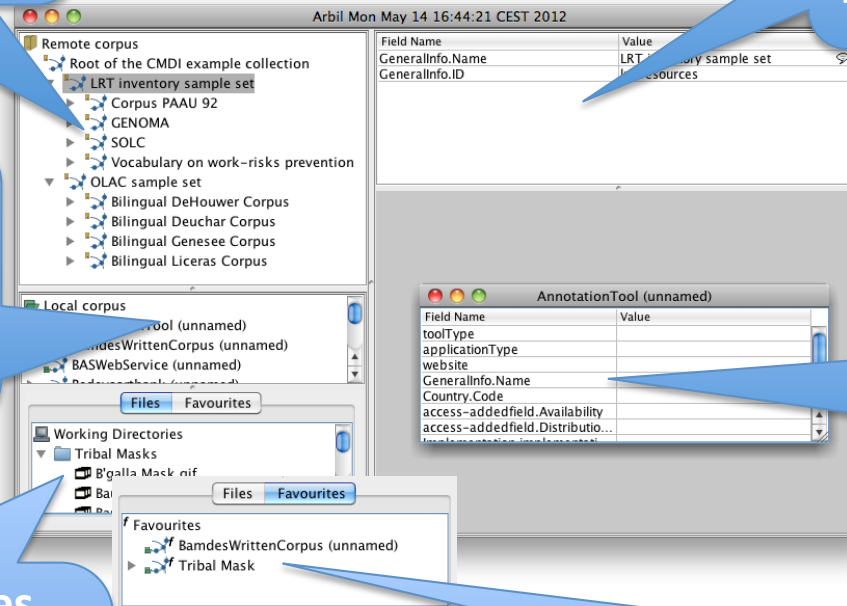
Preview Table (optional)
The currently selected
metadata.

Local Corpus
All newly created
metadata will be
created here

Main Work Area
Multiple tables of
metadata can be
viewed and edited

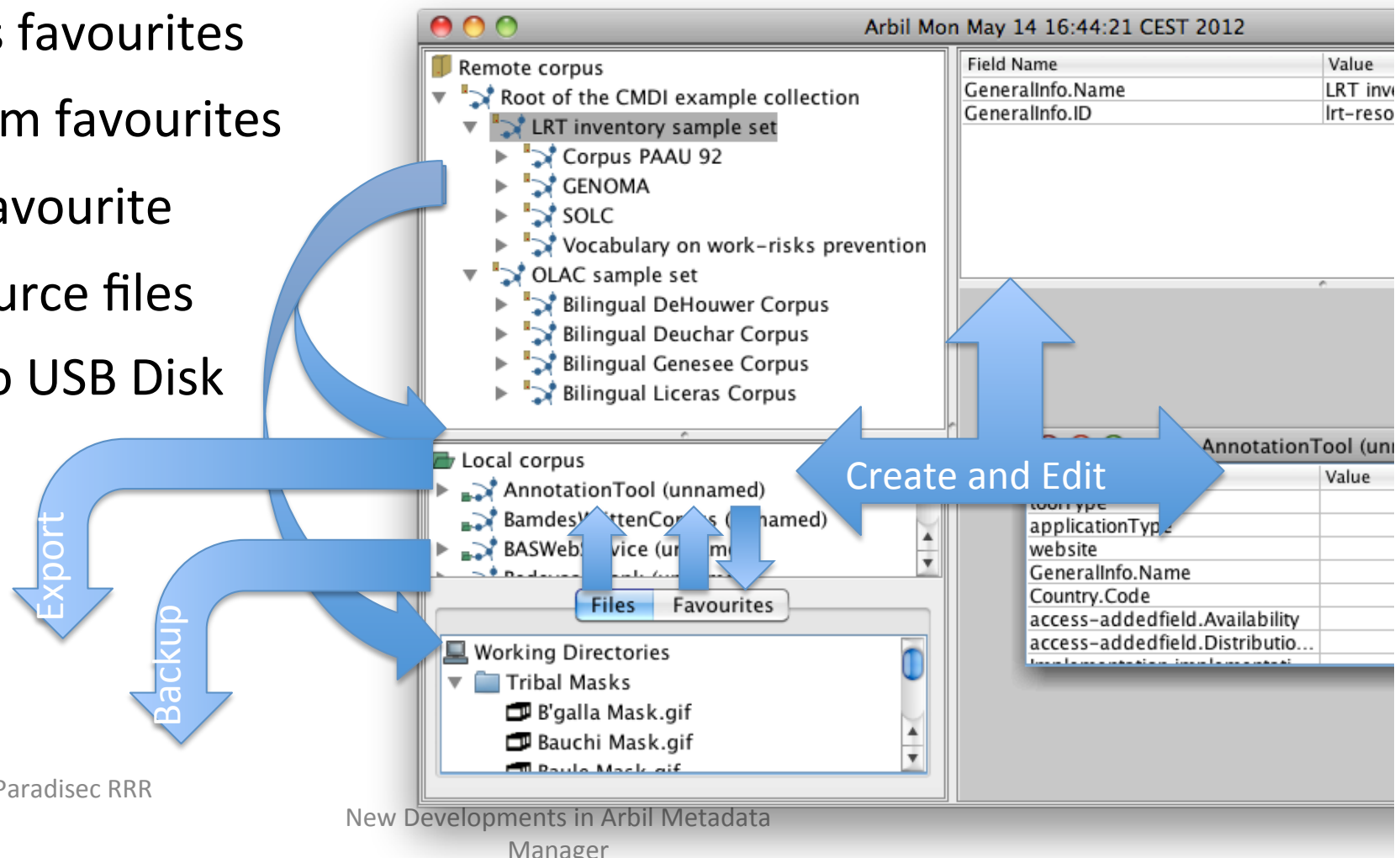
Working Directories
Your data files can be
browsed and associated
with new metadata from
here

Favourites
Frequently used
metadata is saved here
for easy replication



- Create and edit metadata
- Import for offline use
- Import as favourites
- Insert from favourites
- Save as favourite
- Add resource files
- Backup to USB Disk
- Export

Workflow



Metadata Display

- The metadata is displayed in tables and trees, which allow an overview of the metadata and the ability to populate and update many metadata sections in bulk.

Remote corpus

- Root of the CMDI example collection
 - LRT inventory sample set
 - Corpus PAAU 92
 - GENOMA
 - ISO639 (2)
 - Catalan; Valencian
 - Spanish; Castilian
 - LrtCollectionDetails (unnamed)
 - Spain

Metadata files are shown in the tree

Multiple metadata files or subsections can be viewed in a single table

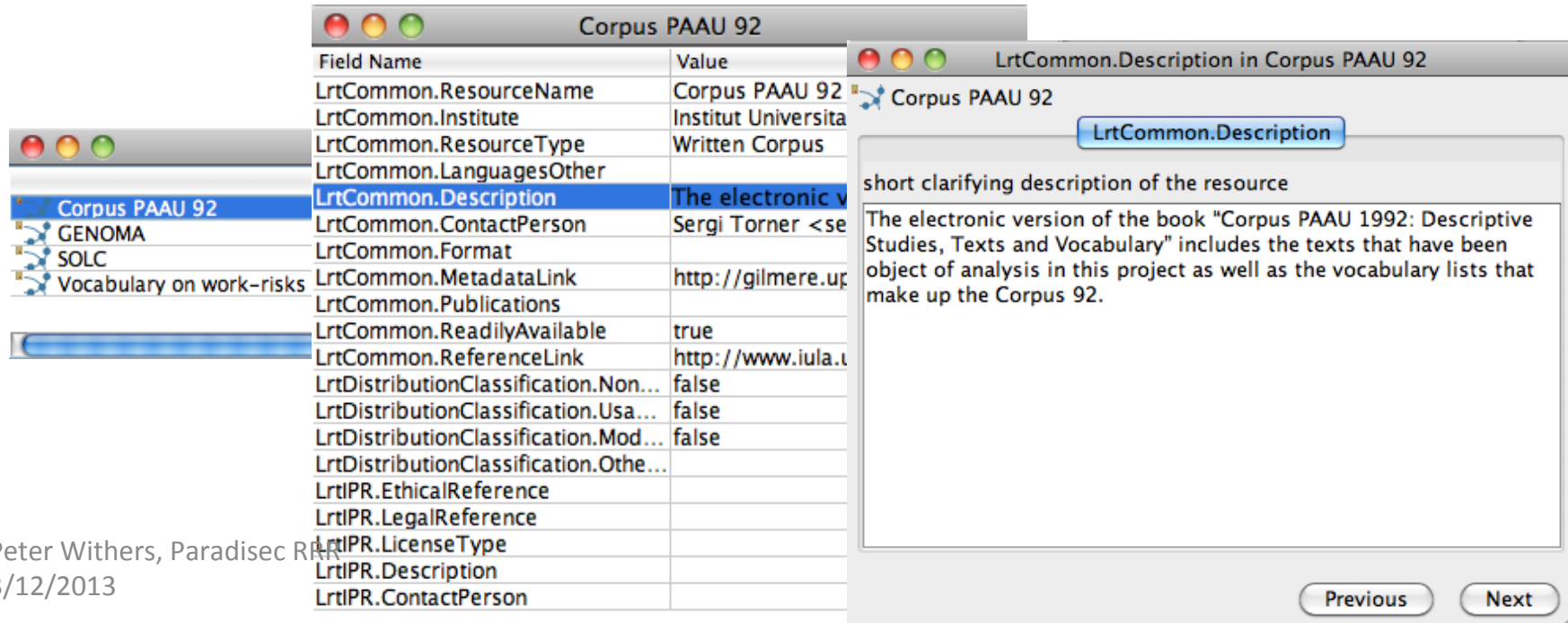
metadata will be highlighted in red

metadata are shown as separate subnodes

				LrtCo...	LrtCommon.Des
Corpus PAAU 92	Corpus PAAU 92	Institut Universitari de Li...		(CV)	The electronic v
GENOMA	GENOMA	Institut Universitari de Li...	Written Corpus	(CV)	Bilingual written
SOLC	SOLC	Institut Universitari de Li...	Application / Tool	(CV)	An orthologic se
Vocabulary on work-risks...	Vocabulary on work-risks prevent...	Institut Universitari de Li...	Terminological Resou...	(CV)	An electronic ve

Table Views

- When multiple metadata files or subsections are viewed in a single table, they are each shown in a separate row.
- When a table shows a single metadata file or subsection, it is shown in the long view.
- Individual fields can be edited in the long field view, which allows each field to be edited sequentially.



The screenshot displays two windows from the LrtCommon interface. The left window, titled 'Corpus PAAU 92', shows a table with the following data:

Field Name	Value
LrtCommon.ResourceName	Corpus PAAU 92
LrtCommon.Institute	Institut Universita
LrtCommon.ResourceType	Written Corpus
LrtCommon.LanguagesOther	
LrtCommon.Description	The electronic v
LrtCommon.ContactPerson	Sergi Torner <se
LrtCommon.Format	
LrtCommon.MetadataLink	http://gilmere.up
LrtCommon.Publications	
LrtCommon.ReadilyAvailable	true
LrtCommon.ReferenceLink	http://www.iula.u
LrtDistributionClassification.Non...	false
LrtDistributionClassification.Usa...	false
LrtDistributionClassification.Mod...	false
LrtDistributionClassification.Othe...	
LrtIPR.EthicalReference	
LrtIPR.LegalReference	
LrtIPR.LicenseType	
LrtIPR.Description	
LrtIPR.ContactPerson	

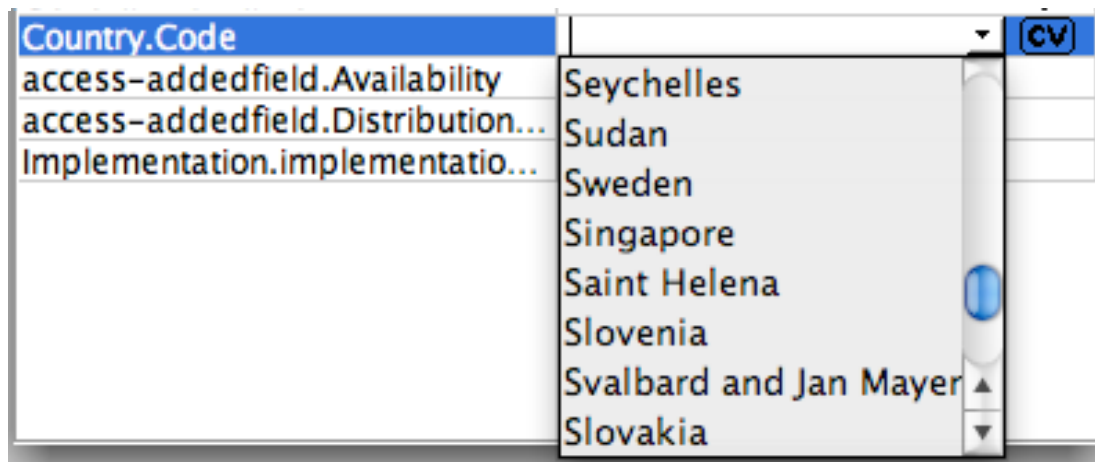
The right window, titled 'LrtCommon.Description in Corpus PAAU 92', shows a detailed view of the 'LrtCommon.Description' field. It includes a 'short clarifying description of the resource' and the following text:

The electronic version of the book "Corpus PAAU 1992: Descriptive Studies, Texts and Vocabulary" includes the texts that have been object of analysis in this project as well as the vocabulary lists that make up the Corpus 92.

Navigation buttons 'Previous' and 'Next' are visible at the bottom right of the right window.

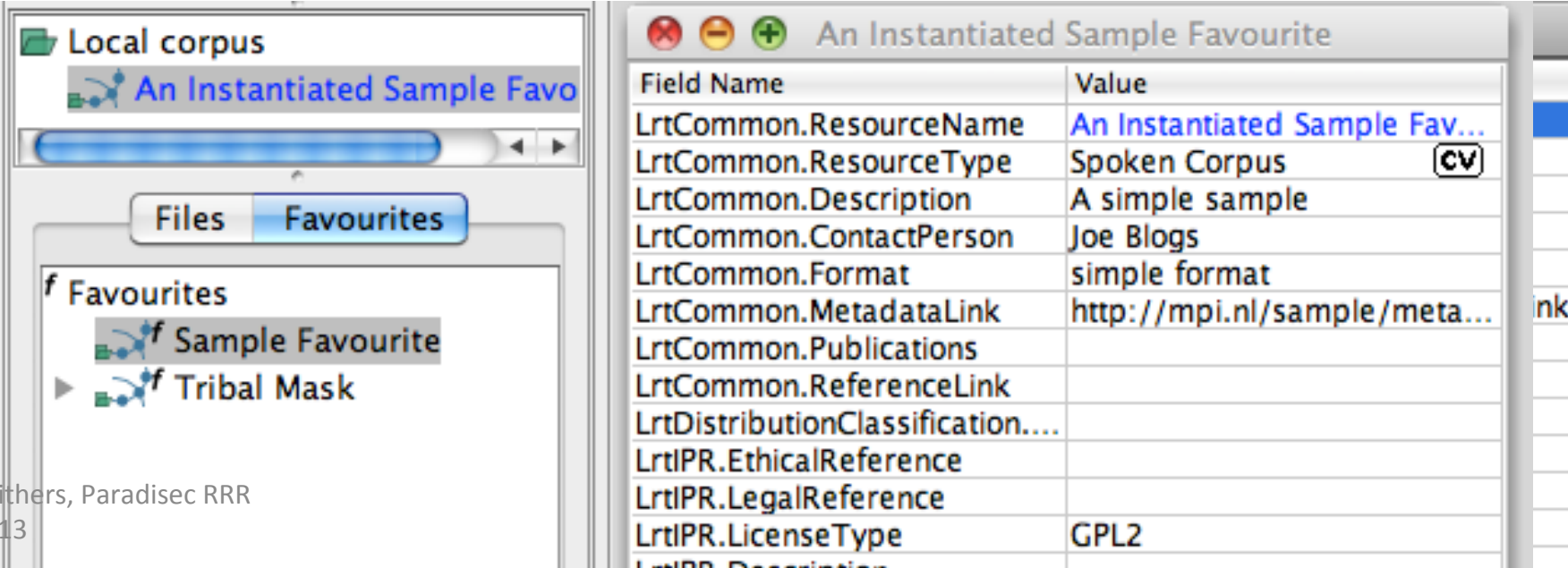
Controlled Vocabularies

- Clarin vocabularies are read from the schema file.
- IMDI vocabularies are read from an XML file.
- Both are provided as dropdown lists in the table when editing.



Using Favourites

- Favourites are snippets of prefilled metadata instances
- New metadata instances can be created from them, leaving only the specific details to be edited
- Either the entire favourite can be used or just the desired sections added to existing metadata

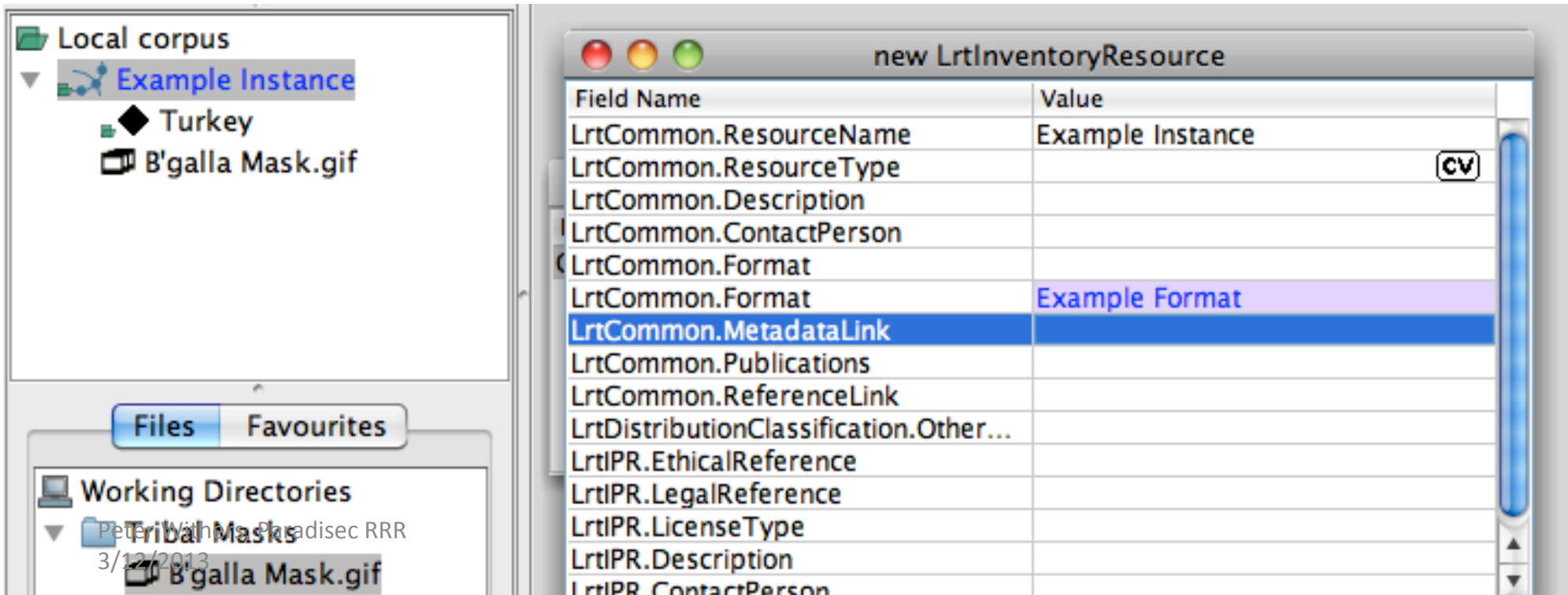


The screenshot shows two windows from a software application. The left window, titled 'Local corpus', has a 'Favourites' tab selected, showing a list of favourites including 'Sample Favourite' and 'Tribal Mask'. The right window, titled 'An Instantiated Sample Favourite', displays a table of metadata fields and their values.

Field Name	Value
LrtCommon.ResourceName	An Instantiated Sample Fav...
LrtCommon.ResourceType	Spoken Corpus (CV)
LrtCommon.Description	A simple sample
LrtCommon.ContactPerson	Joe Blogs
LrtCommon.Format	simple format
LrtCommon.MetadataLink	http://mpi.nl/sample/meta...
LrtCommon.Publications	
LrtCommon.ReferenceLink	
LrtDistributionClassification...	
LrtIPR.EthicalReference	
LrtIPR.LegalReference	
LrtIPR.LicenseType	GPL2
LrtIPR.Description	

Creating Metadata for a Resource

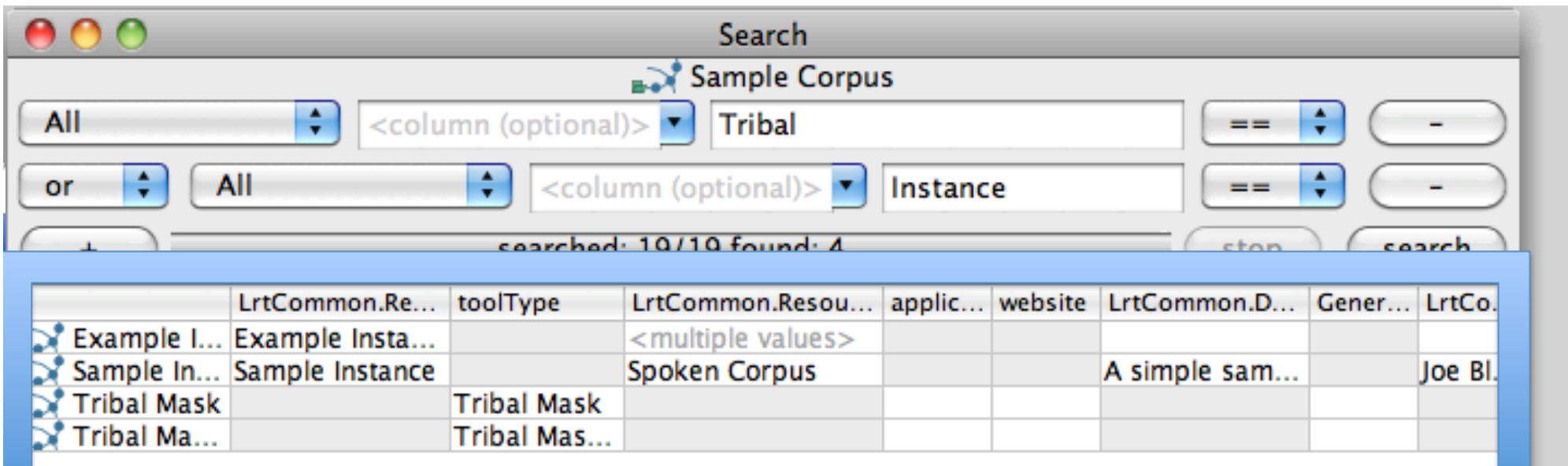
- Create a metadata instance
- Add a data file
- Add a metadata subsection
- Add a field



Field Name	Value
LrtCommon.ResourceName	Example Instance
LrtCommon.ResourceType	
LrtCommon.Description	
LrtCommon.ContactPerson	
LrtCommon.Format	
LrtCommon.Format	Example Format
LrtCommon.MetadataLink	
LrtCommon.Publications	
LrtCommon.ReferenceLink	
LrtDistributionClassification.Other...	
LrtIPR.EthicalReference	
LrtIPR.LegalReference	
LrtIPR.LicenseType	
LrtIPR.Description	
LrtIPR.ContactPerson	

Searching the Metadata

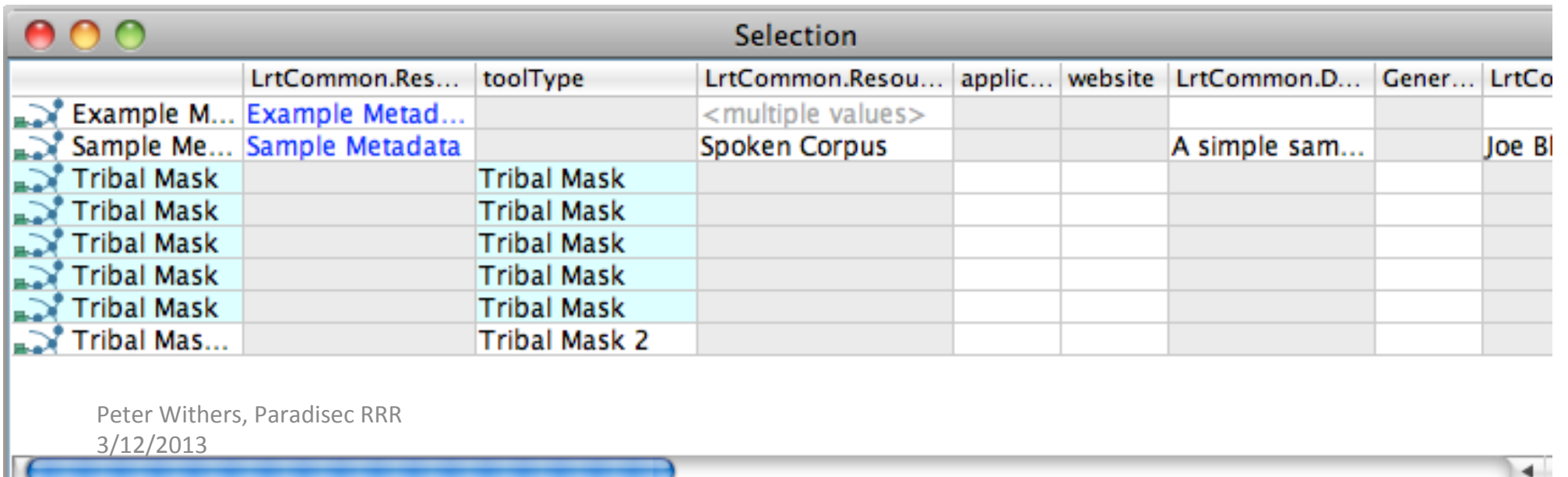
- Searches in Arbil are tree based.
- Multiple search parameters can be entered.
- The search results are shown in a standard table with the usual editing facilities.



	LrtCommon.Re...	toolType	LrtCommon.Resou...	applic...	website	LrtCommon.D...	Gener...	LrtCo.
Example I...	Example Insta...		<multiple values>					
Sample In...	Sample Instance		Spoken Corpus			A simple sam...		Joe Bl.
Tribal Mask		Tribal Mask						
Tribal Ma...		Tribal Mas...						

Find / Replace and Highlighting Cells

- The table can be searched for specific text.
- Selected cells can have the found text substituted.
- Matching table cells can also be highlighted.



	LrtCommon.Res...	toolType	LrtCommon.Resou...	applic...	website	LrtCommon.D...	Gener...	LrtCo
Example M...	Example Metad...		<multiple values>					
Sample Me...	Sample Metadata		Spoken Corpus			A simple sam...		Joe Bl
Tribal Mask		Tribal Mask						
Tribal Mask		Tribal Mask						
Tribal Mask		Tribal Mask						
Tribal Mask		Tribal Mask						
Tribal Mask		Tribal Mask						
Tribal Mas...		Tribal Mask 2						

Peter Withers, Paradisec RRR
3/12/2013

Installing Arbil

- There is a link to ARBIL on the MPI website
<http://tla.mpi.nl/tools/tla-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).
- The manual and user guide are also available for download on the same page



The screenshot shows the Arbil page on the The Language Archive website. The page header includes the Max Planck Institute for Psycholinguistics logo and the Language Archive logo. A navigation menu is visible with links for Home, Team, Projects, Tools, Resources, Events, Forums, and Contact. The main content area is titled "Arbil" and describes it as a general metadata editor, browser & organizer tool for IMDI, CMDI and similar metadata formats. It provides links for "More information...", "Release history...", "System requirements...", and "How to cite Arbil...". A "Download" section offers three options: "Run via webstart" (with a green arrow icon), "Download the Windows installer" (with a Windows logo icon), "Download the Debian Linux package" (with a Debian logo icon), and "Download the Mac installer" (with a Mac logo icon). A "Screenshot" of the Arbil interface is shown to the right. The footer includes contact information for The Language Archive, Max Planck Institute for Psycholinguistics, and social media links for Facebook, Twitter, and RSS.

Arbil in the Future

- MArbil
 - Simplified interface
 - Less workflow specific
- YAAS
 - Online search
 - Potential mobile application

MArbil

- Highly simplified interface
- Minimal workflow requirements
- The local cache will not be relevant
- The local corpus can be any directory/s
- The metadata files reside next to the data file
- The OS's file browser is the main entry point
- Search widget:
 - Immediate use of the metadata
 - Gathers the data/metadata into one interface

YAAS prototype (yet another archive search)

CGN-2013-11-013-A

Available Documents: 12894 Missing Documents: 0

remove Session (12767) Date (12767) equals 1997

remove All Types (12894) Name (83975) equals Dutch

add search term

Inter Dutch
Dutch regions
interviews with teachers of Dutch

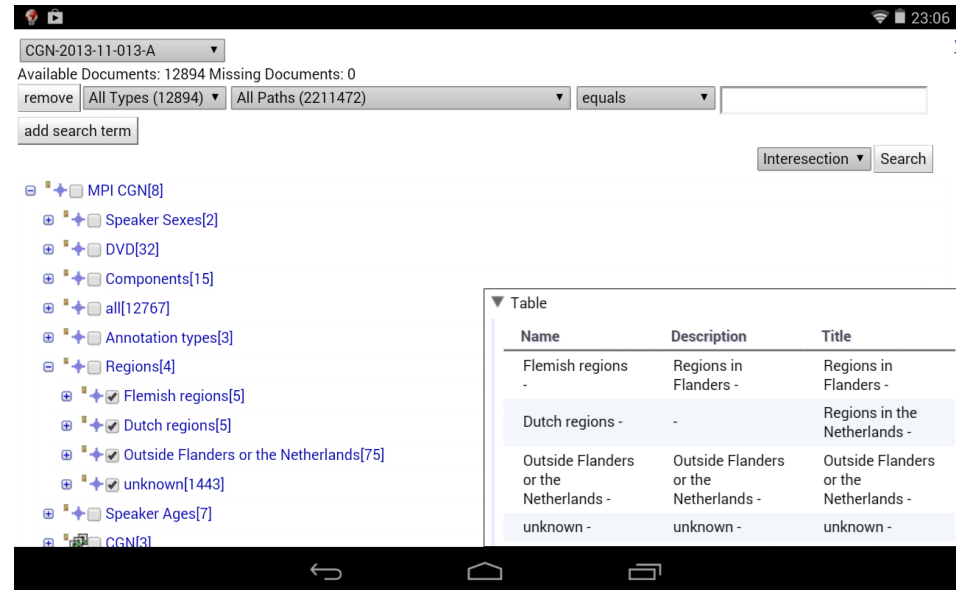
Search Results: intersect (Name equals Dutch) (Date equals 1997) [42]

- ⊕ fv800191[6]
- ⊖ fv800334[6]
 - ⊕ Actors (2)[2]
 - ⊕ WrittenResources (4)[4]
 - ⊕ MediaFiles (2)[2]
 - ⊖ Content (Discourse)[1]
 - ⊖ Languages (1)[1]
 - ◆ Dutch[-1]

x	Metadata Link
Name	Dutch
Id	RFC1766:x-sil-DUT
 - ⊕ Sources (1)[1]
 - ◆ Project[-1]
- ⊕ fn007333[6]
- ⊕ Peter Withers Paradisec RRR
3/12/2013
- ⊕ fv800051[6]
- ⊕ fv800287[6]

YAAS prototype (yet another archive search)

- XML database like KinOath
- Rendered in HTML5
- Can also be compiled into mobile application
- Potentially a cloud based design in the future



CGN-2013-11-013-A

Available Documents: 12894 Missing Documents: 0

remove All Types (12894) All Paths (2211472) equals

add search term

Intersection Search

- MPi CGN[8]
- Speaker Sexes[2]
- DVD[32]
- Components[15]
- all[12767]
- Annotation types[3]
- Regions[4]
 - Flemish regions[5]
 - Dutch regions[5]
 - Outside Flanders or the Netherlands[75]
- Speaker Ages[7]
- CGN[3]

Name	Description	Title
Flemish regions -	Regions in Flanders -	Regions in Flanders -
Dutch regions -	-	Regions in the Netherlands -
Outside Flanders or the Netherlands -	Outside Flanders or the Netherlands -	Outside Flanders or the Netherlands -
unknown -	unknown -	unknown -

← → ↻ lux17.mpi.nl/ds/yaas2/ ☆

YAAS Web Prototype

Translations of Arbil

- Arbil has been translated into a few languages
 - English
 - Spanish
 - Italian
 - German
- These languages are known to our current student assistants
 - Other languages will be added when possible
 - As the student assistants change our ability to support a given language will also change
- Launchpad (or other)
 - Provides a translation tool
 - Facilitates community based translations
 - Any language with sufficient community interest can therefore be maintained

Conclusion

- Arbil has been developed with a strong focus on the workflow of the DoBeS community.
- Future work is aimed at a wider community.
- User can view and edit metadata without mandating any sequence of entry.
- Warnings will be shown, if the metadata does not comply with the requirements.
- It is hoped that the features of Arbil will lead towards the recording of metadata at an earlier stage resulting in greater detail and better quality of that metadata.