Metadata Management with Arbil

Peter Withers

Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands
peter.withers@mpi.nl

Abstract

Arbil is an application for creating and managing metadata for research data such as audio, video or textual data. 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. A number of metadata formats are supported and Arbil has been designed as a local application so that it can also be used offline, for instance in remote field sites. The user can view and edit the metadata in tables in the order that the information is available, if the metadata does not comply with the requirements the user will be warned but will not be prevented from entering it in the meantime. It is hoped that the features of the application will lead towards the recording of metadata at an earlier stage resulting in greater detail and better quality of that metadata. If this improvement in workflow is achieved then the metadata will be entered sooner and reassessed during the research process, which will greatly improve the quality of that metadata.

Keywords: Metadata, Editor, Resources, Corpus, Linguistics, IMDI, Clarin, XML, Schema, Archiving

1. Introduction

Arbil is an application designed to create and manage metadata for research data and to arrange this data into a structure appropriate for archiving. 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. A number of metadata formats are supported and Arbil has been designed as a local application so that it can also be used offline, for instance in remote field sites. The metadata can be entered in any order or at any stage during the process and then exported with the data files for use in the archive or as a backup of the current work. Once the metadata and its data are ready for archiving and an Internet connection is available it can be exported from Arbil and in the case of IMDI it can then be transferred to the main archive via LAMUS (Broeder et al., 2006) (archive management and upload system). In this paper we discuss why the use of a dedicated metadata editor is of benefit and why Arbil was written, we also discuss how this application can be used to create and edit metadata.

2. Why use a metadata editor

There are many reasons to provide metadata, yet if the process of creating and managing that metadata is difficult, the quality and completeness will suffer. It is a reasonable assumption that from the point of view of the researcher collecting the primary data, that this data is considered valuable and worth preserving with metadata so that it can be subsequently found, understood and referenced in future publications. In many cases there can also be an obligation to provide to the speakers of the language being researched and their descendants access to the collected material, and this would not be complete without metadata. From the point of view of the archivist, the task is not just to preserve the data, but also to organise the material in a structured way such that it can be identified, searched for and accessed when required. For these reasons it is important that we provide a tool that makes the process of creating metadata simple and transparent, reducing repetitive tasks whenever possible.

A metadata tool must have at very least all the functions that a basic text editor provides, such as copy, paste, find and undo. A simple text editor at first glance has the advantage of being simple to use and very flexible. However, it does not enforce any structure on to the metadata being edited. Conversely if a structured metadata editor is confusing, or not reliable, or does not have the basic set of functionality to which the user is accustomed, then the users may end up resorting to an unstructured tool instead. Which can lead to inconsistency of the metadata produced, hence it is crucial that an easy to use and reliable tool is available for the task. Arbil is designed to fill this need by providing an intuitive modern interface in which to create and manage metadata for the data files being archived. Features like drag and drop are used extensively both for constructing a hierarchical corpus tree structure and for adding nodes to tables for viewing and editing. Bulk editing of metadata can be done for instance via copy and paste, which allows a string of text to be pasted into multiple fields of multiple rows, or to paste multiple fields into the matching fields of multiple rows. Whenever a field is edited the changes are stored in an undo / redo buffer which allows all the changes made since the last save to be undone or redone. Arbil supports both IMDI (Broeder and Wittenburg, 2006) and Clarin (Váradi et al., 2008) formats and through the use of XML schema files additional formats can potentially be supported. Both the IMDI and Clarin formats allow the metadata to be arranged into corpus tree structures. Arbil displays trees of metadata in its user interface 'remote corpus', 'local corpus', 'favourites' and the directories containing the data files. Snippets of frequently used sections of metadata can be collected in the favourites and then easily utilised in the process of constructing new metadata, greatly reducing the amount of repetitive data entry.

3. Why Arbil was created

Arbil came into existence as a result of a meeting between members of the DOBES (Wittenburg et al., 2002) commu-

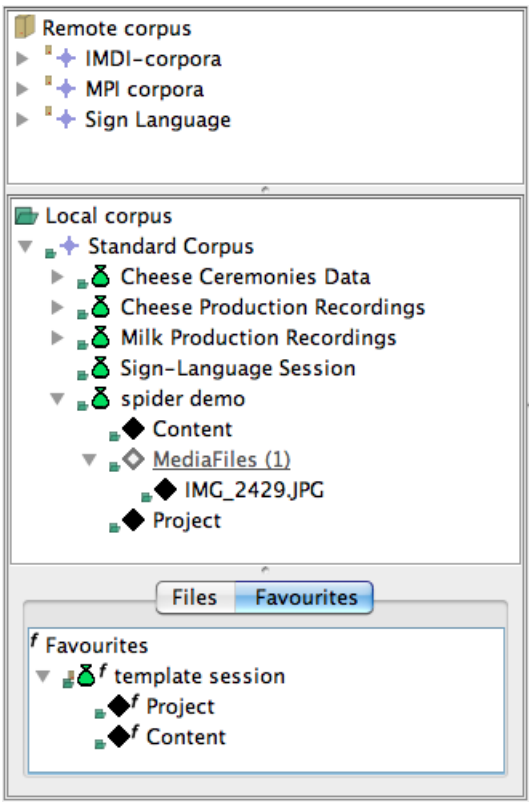

Figure 1: Tree display of the metadata.

nity and members of the MPI developers team where the need was recognised for an offline metadata editor with tabular functionality, something the previous IMDI Editor lacked. A prototype application called Linorg was developed by the author, and based on the feedback given during discussion with members of the DOBES community; from that prototype he subsequently developed what is now Arbil. The development of Arbil has also benefited from discussions with many linguists at the MPI and the experience gained from the previous metadata applications developed at the MPI. Arbil contains many features in order to fulfil the wide-ranging needs expressed while maintaining the functionality of the previous IMDI Editor tool. While Arbil is primarily designed to create metadata, it also has functions to help organise the collected material and create a local well-organised corpus before it is archived. These functions include the ability to search for and compare metadata, and to search for and open the data files in the relevant application. Arbil continues to be actively developed to extend these features further.

Often researchers are working in a field site where there is limited or no Internet connection. For this reason it is important that a tool such as Arbil is able to work correctly when offline. Arbil achieves this by keeping a local copy of all the required files such as controlled vocabularies and will update them if required from the server when an Internet connection is available. One of the most network intensive activities is browsing the remote archive; clearly this will not be possible without a network. However, for this reason, Arbil has the ability to mirror branches from the main archive so that they can still be referred to offline and in the field.

Figure 2: Metadata node view.

4. Entering metadata

Some metadata editors, for instance the IMDI Editor, requires that the user enters the metadata in a predefined order making it impossible to move forward until a value is entered. While this is useful when the data to be entered is minimal and or the required information is completely available at the time of entry, in reality this is likely to result in a situation when the data is not fully available and the user is forced to either fragment the metadata by recording some of it outside the system or by entering dummy metadata with the good intention of fixing it later. Both of these workarounds can lead to inconsistency of the metadata recorded. This issue is addressed in Arbil by allowing the metadata fields to be completed when the information is available and to simply warn a user when something is missing or is not in the required format. At the point of exporting the metadata, all files are checked for inconsistencies and warnings are given if there are issues. Only at the point of pushing the metadata into the archive will the user be blocked if they have not correctly completed all the required fields.

In Arbil the metadata is viewed in tables, which can contain a single node of metadata as a list of fields, or many different nodes, each with its fields as a separate row in the table, or all the nodes of one metadata file inline. This tabular view of the data allows multiple metadata nodes to be quickly viewed across the rows of the table. The metadata can be can edited in any table in which it is viewed, for instance in the search results table or a table of individually selected metdata nodes. These manually constructed tables can be assembled by selecting metadata nodes of interest and drag-and-dropping them into a table.

In many metadata sets the number of fields required to describe the data and its context can be extensive; this can make it difficult for a user to see their relevant information at a glance. In order to accommodate this the table columns in Arbil are customisable, so that only those relevant to a particular user need be displayed. These selected sets of columns viewed in a table can be saved and then easily applied to any table, and if required, a default combination of columns can be selected so that new tables show only the required information. In order to further visualise the metadata in the table, the columns can be resized or sorted

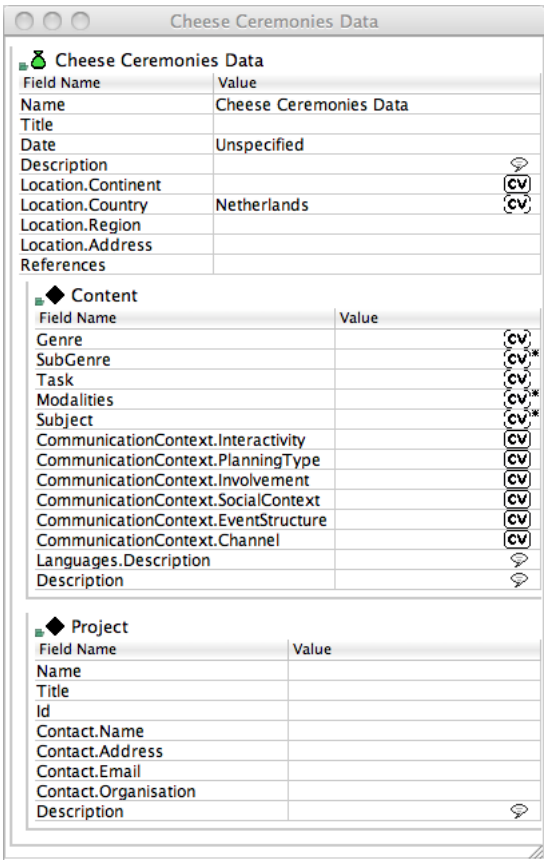

Figure 3: Metadata file view.

on any column and reordered. Rows can easily be added and then dragged from one table to another, and the cells can be highlighted based on matching text. Because much of the metadata is hierarchical with multiple sub nodes in a single file, this cannot always be displayed in a single row of a table. For instance in the IMDI metadata format actors, written resources and media files are sub nodes within a single session file. However, in this case additional columns can be displayed where the name and icons of the sub elements are displayed in a single cell of the row.

When there are many fields to fill in for a given metadata set, it is important to clearly see what each fields is intended for and which fields are of a higher priority than others. For this reason a description can be provided (in the metadata format specification) for each field explaining the intended usage and this is displayed in the tooltip of that field. When a field is set as mandatory it will be given a colour highlight if the metadata is not filled in. Likewise in the case of fields requiring specific formatting, such as date fields, the text will be highlighted when the formatting is incorrect.

Creating and editing of the metadata is only one part of a much larger workflow, hence it is necessary to both import and export this metadata in Arbil. Any valid IMDI or CMDI metadata file can be imported into Arbil. The data files that the metadata describes can optionally be imported at the same time, for instance when migrating or merging from one computer to another. If a backup is required then all the metadata and data files within Arbil can be exported into a self-contained directory, for instance onto a USB hard

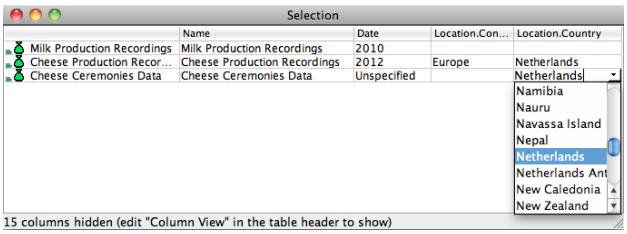

Figure 4: Multiple metadata files view.

drive. In the case of IMDI the resulting export can then be uploaded into LAMUS and from there inserted in the archive. During both the import and export processes, all of the metadata is checked for errors and a list of warnings given if any are found. The metadata in Arbil can be exported in other formats via XSLT transforms and one such transform is provided with the application that converts from IMDI into HTML. In addition, when any metadata is displayed in a table the contents can be copied and then pasted into a text editor or into spreadsheet.

5. Conclusion

Arbil has been developed with a strong focus on workflow and usability. It allows the user to view and edit the metadata in tables without mandating any particular order of metadata entry while warning if the metadata does not comply with the requirements. It is hoped that the features of the application will lead towards the recording of metadata at an earlier stage resulting in greater detail and better quality of that metadata. It is also hoped that this metadata will prove useful for the linguists during the process of their research. Creating metadata at the time the data is collected can assist workflow by helping to keep track of the collected data files. By providing a way to organise these data files and utilise the metadata for searching the collected data and to backup the current data with its metadata, it is hoped that Arbil will assist the workflow of the researcher. If this improvement in workflow is achieved then the metadata will be entered sooner and reassessed during the research process, which will greatly improve the quality of that metadata. Hence, if the chore of entering of metadata at the end of a project is replaced by useful metadata throughout the life of the project it is likely to be of benefit to the process as a whole.

6. References

D. Broeder and P. Wittenburg. 2006. The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies*, 1(2):119–132.

D. Broeder, A. Claus, F. Offenga, R. Skiba, P. Trilsbeek, and P. Wittenburg. 2006. LAMUS : the Language Archive Management and Upload System. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 2291–2294, Genoa. European Language Resources Association (ELRA). www.lat-mpi.eu/papers/papers-2006/lamus-paper-final2.pdf.

T. Váradi, S. Krauwer, P. Wittenburg, M. Wynne, and K. Koskenniemi. 2008. Clarin: Common language resources and technology infrastructure. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1244–1248, Marrakech. European Language Resources Association (ELRA). www.lrec-conf.org/proceedings/lrec2008/pdf/317_paper.pdf.

P. Wittenburg, U. Mosel, and A. Dwyer. 2002. Methods of Language Documentation in the DOBES project. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 36–42, Las Palmas. www.lrec-conf.org/proceedings/lrec2002/pdf/221.pdf.