LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 8 (2014), pp. 307-314
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24609

Arbil: Free Tool for Creating, Editing, and Searching Metadata

From http://tla.mpi.nl/tools/tla-tools/arbil/

Reviewed by Rebecca Defina[1], *Max Planck Institute for Psycholinguistics Nijmegen*

1. OVERVIEW. Arbil[2] (Withers 2012) is a free Java-based software program for creating and editing metadata. It was created by Peter Withers, of The Language Archive at the Max Planck Institute for Psycholinguistics, for use within the DoBeS[3] (Dokumentation bedrohter Sprachen, 'Documentation of Endangered Languages') program. Arbil is now being maintained by a team led by Peter Withers and Twan Goosen, who are developing and extending it for a wider user group.

The creation of metadata is an important part of any project that collects data. In this review, I will focus on language description and documentation style projects and the kinds of linguistic and other data that they tend to collect. Metadata is information that describes the content of other data. Metadata usually describes the type of data, for instance whether it is a video recording, a written text, or a photo, and the format of the data, for instance MPEG1 or MPEG2. It should also describe when and where the data was collected. There should also be a good description of all the people involved. For instance, if you are describing a transcription you need to know who is speaking in the recording, who made the recording, and who did the transcription. For more information about linguistic metadata see Bowern (2008:56-59) and Thieberger & Berez (2012). Metadata makes it possible for community members and other researchers to find out what kind of data you have and to find material they are looking for. Metadata is also necessary for yourself as a researcher, to keep a lasting record of the details of your data and allow you to locate recordings and information for many years to come. I am very glad that I have good metadata for recordings I made five years ago. Without it I would be completely lost, looking for elicitation sessions on a particular topic or the name of the woman who told that story about the dog. Arbil is not only a good tool for the initial creation of metadata, it can be used later to search your metadata and open the associated files directly.

There are several different formats for linguistic metadata in use. Arbil is intended for the creation of metadata in either the ISLE MetaData Initiative (IMDI)[4] or the Component MetaData Infrastructure (CMDI)[5] format. Both formats are XML-based templates for metadata. They provide a list of fields, some of which are obligatory and must be filled in in order to have a completed metadata file, and others which are optional. This both orders and constrains the type of metadata created. This constraint on the structure of the

[1] I would like to thank Felix K. Ameka, Saskia van Putten, and Peter Withers for their comments and useful discussion.

[2] http://tla.mpi.nl/tools/tla-tools/arbil

[3] http://dobes.mpi.nl

[4] http://www.mpi.nl/IMDI

[5] https://www.clarin.eu/content/component-metadata

E-ISSN 1934-5275

metadata can be a very good thing as it ensures consistency, which is otherwise hard to achieve without a template. The template structure also allows the metadata to be automatically interpreted so that it can be archived, searched, and compared. The IMDI format has been around for over a decade and was designed by the DoBeS team for use with material collected during language documentation. The CMDI format is newer and was developed by Clarin[6], an initiative for integrating language resources and tools across Europe. The CMDI format is intended for use by any researcher in the humanities or social science, not only for linguistic fieldworkers. The main difference between the two formats is that the IMDI format is fixed: it comes with the fields and parameters already defined; the CMDI format is much more customizable by the user. This customizability makes it usable by a wider group of researchers and allows individuals more flexibility. It is also possible to configure Arbil to create metadata in other XML formats via the use of custom templates or schema files.

Arbil is currently the best general editor of IMDI or CMDI format metadata for most linguists. It is possible to use XML editors such as Oxygen[7] to write your own IMDI or CMDI format metadata; however these require fairly advanced knowledge of XML. There are also some other specialized tools for writing IMDI or CMDI metadata. These are often available online only or specialized for particular purposes and thus have a limited functionality compared to Arbil. For example, ProFormA[8] (Dima et al. 2012) is an online tool designed for projects within the Center for Sustainability of Linguistic Data at the University of Tübingen. It provides a very easy way to enter metadata for some types of data and upload them to the ProFormA server, but is currently limited to a small set of data types. There is also the IMDI Editor[9], which Arbil was built to replace. This tool can still be used to create IMDI format metadata, however it is quite similar in style to Arbil, and Arbil has massively improved the interface and functionality, for instance allowing editing of multiple files.

Arbil was first released in 2008 and I have been using it since 2009. It has been actively developed and improved over this time following user comments and requests. This review is based on version 2.4.36550. There is good support available for Arbil. The manual (Saad et al. 2012) is very detailed and helpful and is available both from the Arbil website and via the program's help section. There is also good support available in the online support forum, which can also be accessed via the website or directly from the program. A list of currently known issues and requests is available via the website, and most issues include a reference to a version number giving an indication of when they are expected to be resolved.

2. GETTING STARTED. Arbil can be used either through a web browser or installed on your computer (Windows XP or higher, Mac OS X, or Debian[10]). If you already have Java installed, the quickest and easiest way to start using it is via a web browser. I would recommend doing this as an excellent way to test out Arbil. For long-term, regular use, I prefer

[6] https://www.clarin.eu

[7] http://www.oxygenxml.com

[8] http://www.sfs.uni-tuebingen.de/nalida/proforma/web

[9] http://tla.mpi.nl/tools/tla-tools/older-tools/imdi-editor

[10] https://www.debian.org/

to install a stand-alone version of Arbil. Your choice will, of course, vary according to your preferred working style and needs.

Arbil is designed to create metadata in the IMDI or CMDI format. When you first start up Arbil, you will have to choose which format you wish to use. This will often be determined by the requirements of the archive you are working with. If the archive does not require metadata in a particular format, you are free to choose. All IMDI metadata are compatible with, and can be converted to, CMDI, but CMDI metadata is generally not IMDI-compatible.

Arbil is structured around Remote and Local Corpora, shown in the top and middle left of the screen respectively, see Figure 1. The Remote Corpus is a connection to archives using IMDI or CMDI format metadata. By default, it links to the data in The Language Archive. Other IMDI/CMDI-compliant archives can also be added. This link allows you to view and download metadata and the associated data (if accessible) from the archive. The Local Corpus is your own data that you are working on. The Remote Corpus is most useful when you have already archived data with the archive. You can then download your archived metadata and associated data files from the Remote Corpus to the Local Corpus to edit the metadata, search it, or add any new material. Access to the Remote Corpus is the only Arbil feature which is not available when working offline without an internet connection.

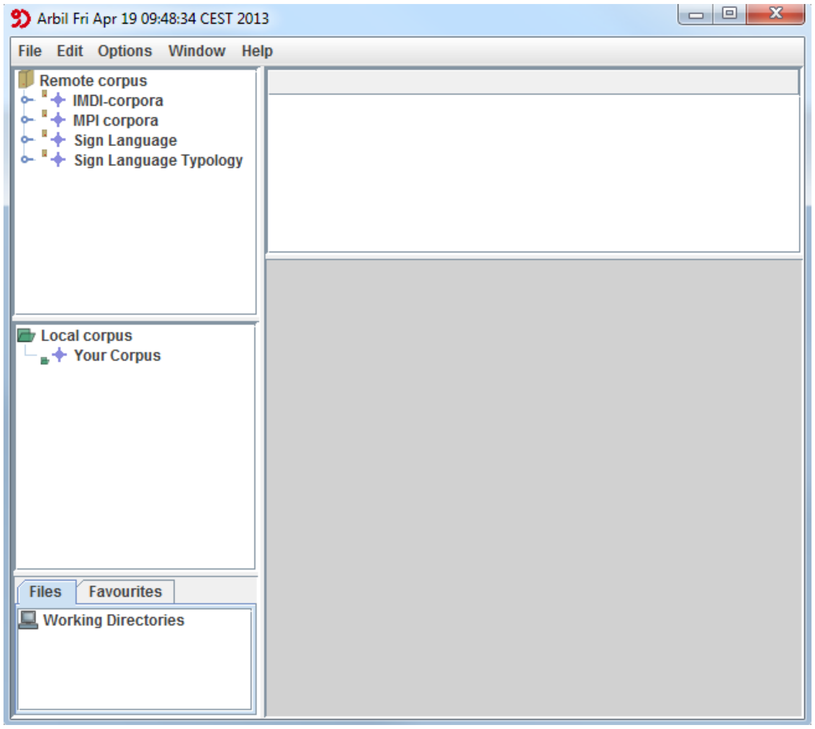
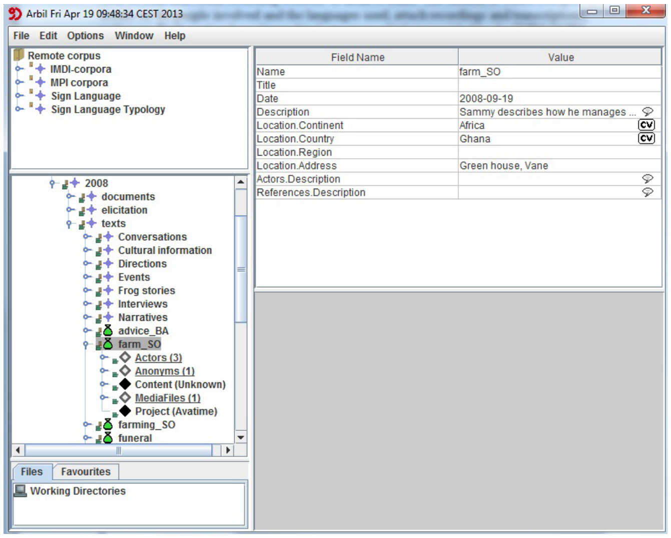


FIGURE 1. Screen shot of Arbil showing the Remote Corpus at the top left, the Local Corpus in the middle left, and the Files and Favourites section at the bottom left.

Arbil is primarily designed for use by individuals. It is, however, possible to set it up so that multiple users can work with and edit the same Local Corpus, provided they all have access to a shared drive. In the Options tab, under Local Corpus Storage Directory, there is an option to change where the Local Corpus is stored. If you select a location that all users have access to, and make this change for all users, then you will be able to work with the same Local Corpus without needing to update the archive and download it every time you make changes. The only caution here is that multiple users should not work with Arbil at the same time.

3. CREATING METADATA. Arbil creates a corpus of data sessions, each described in a Session Node. Exactly what counts as a data session is flexible. Most often it will be one of your recording sessions, and the Session Node will contain all data files associated with that session, for instance any video or audio recordings made, photos taken during the session, and any annotations of these files such as transcriptions or translations, as well as all metadata relating to these files. A Session Node could also refer to a single data file, perhaps a photo that was taken or given to you, or a written text you acquired. Each Session Node has a hierarchical structure, which is provided by the IMDI/CMDI format. For instance, all people involved in the session are grouped together in a node called Actors, and all media files are also grouped together under a single node; see Figure 2.

Arbil also allows you to create a hierarchical structure to organize your individual Session Nodes. This structure is fully customizable. You could create a completely flat corpus or group your sessions into whichever categories you find useful; see Figure 2 for an example from my corpus. You can edit the structure of your corpus at any time. You do not need to decide on an organizational system before you start entering session information, and you can change your system at any time.



Field Name	Value
Name	farm_SO
Title	
Date	2008-09-19
Description	Sammy describes how he manages ...
Location.Continent	Africa
Location.Country	Ghana
Location.Region	
Location.Address	Green house, Vane
Actors.Description	
References.Description	

FIGURE 2. Hierarchical node and corpus structure. The blue ✦ symbol is used for a corpus node that can be expanded. The green sack symbol 💰 is used for individual sessions.

One downside of Arbil is that it takes some time to understand how best to use it. The full session structure is not transparent and it takes some time and practice to learn what information can be placed where. Once you have become familiar with the structure, it works well. A good way to familiarize yourself with the program is to browse other corpora in the Remote Corpus to see how others have organized their data.

Another issue for new or infrequent users is that some fields require data to be entered in a particular format, but Arbil does not yet provide much guidance in what format is required. For instance, dates must be entered as YYYY-MM-DD. This is a good thing as it keeps the format consistent within your own data and also across the archive. The problem is that Arbil does not help you to enter information in the correct format. If you enter data in an incorrect format, the text is colored red. There is no message to tell you that the format is incorrect or what the correct format is. You have to figure out the required format and remember it yourself. This is a known issue, though, and hopefully a message to tell you the correct format will be coming soon in a new version.

Once you become familiar with the structure of Arbil session entries, I find the user interface works very well and it is easy to enter information. The Arbil interface makes it easy to get a quick overview of data sessions and to edit these sessions in a modular way. For instance, it is easy to find and edit information on the people involved in a data session without needing to wade through other information. The structure can at times be repetitive and some fields will feel unnecessary for some data types. This is an issue with the metadata format, rather than with Arbil, per se. This is the downside of using Arbil for a project that does not require IMDI or CMDI format metadata; you must meet the format requirements even though you have no external reason for doing so. Nevertheless, most of these fields can be filled in quickly and easily, and I have found the benefits of using Arbil a small price to pay for unnecessarily following the IMDI format requirements.

One really useful feature of Arbil is the Browse For Resource File method of adding media files and written resources. This allows you to locate the file in your file directory and link it to Arbil. Using this method, Arbil will automatically fill in the name of the file, its location, type, format and size. This saves you a good amount of time and bother. The only problem is that this process can sometimes take a long time, especially if the file is large or the connection to the file location is slow. There is no way to bypass this automatic look up, and Arbil will sometimes freeze while it deals with a large file. If this happens, you just need to wait until the look up is finished before carrying on.

The feature that I find most useful in Arbil is the Favourites. You can add any part of a session to your Favourites directory. For instance, you can add the details of a participant in a recording to the Favourites and then whenever that person participates in another recording you can add this information directly from the Favourites without needing to re-enter it or locate the other recording session in order to copy the information over. You can also add whole Session Nodes to your Favourites directory. In this way, you can create templates for commonly used session types, such as elicitation sessions, conversations, and experiments. You can fill out all the information which is common to sessions of this type and leave blank any fields which will vary, then just save it to your Favourites directory and call it up from there whenever you need to enter a session of that type.

4. EDITING METADATA. You can, of course, also use Arbil to edit metadata you have already entered[11]. This could be a small change to one session node, for instance correcting a typo or adding a new transcription. In other cases, you may decide you need to make major changes. For instance, you may need to change some information about a person who appears in multiple sessions, or you may have decided to replace a term you have used in many sessions with another term. It is these changes affecting many sessions which can be potentially daunting; luckily Arbil provides tools to assist with this. You can use the Edit Multiple Cells feature to enter or edit data in a set of fields across a set of sessions. This is a good method for changing a field for a person who appears in multiple sessions. This can also be used to enter the data in the first place. For instance, you could decide to leave the continent and country fields blank when creating all your sessions and then fill them in all at once using this tool. This tool works well if you need to change an entire field. If you need to change only part of a field, for instance a single word or phrase that occurs within different fields across sessions, the Find And Replace tool can be very useful.

5. SEARCHING METADATA. Arbil is also very useful once you have entered all your metadata. Arbil has excellent search functionality, which is being actively improved. For instance, the ability to search using regular expressions, wildcards, and constraint options is expected to arrive with version 2.6. This makes it easy to find the information and sessions you are looking for. The best feature of using Arbil to search your metadata is that you can open the associated local data files directly from Arbil. So, you do not need to locate the files in your file directory once you have worked out which session contains the information you are after. The only warning here is that, at present, Arbil can only search over about 1000 entries. So if you want to use the Search or the Find And Replace feature, you will need to make sure that there are less than 1000 session nodes in your local corpus.

6. FUTURE DEVELOPMENTS. Arbil is being actively maintained and developed with bug fixes and improvements being released regularly. Version 2.4 has been the longest running version with about one year between updates. However, version 2.5 is due to be released shortly. There are also new tools that connect with Arbil starting to come out. It is already possible to use KinOath[12] to connect your Arbil data to kinship information. The Cologne Language Archive Services[13] will shortly be releasing a tool that automatically generates session nodes in Arbil from a list of recordings. These automatically generated session nodes will contain only quite sparse information, such as the filename and date of recording, but any assistance is appreciated with metadata creation. The Language Archive is also working on an online metadata search tool called YAMS (Yet Another Metadata Search)[14] and a new, reduced, and simplified version of Arbil called Marbil, or 'mini-Arbil.'

6. CONCLUSION. Writing metadata is never a truly enjoyable task; however, it is a necessary evil. While there are still some problems with Arbil, and it can take some time to get

[11] If you are using The Language Archive, you must download a fresh copy of your corpus before you make any changes, otherwise you will not be able to upload your changes.

[12] http://tla.mpi.nl/tools/tla-tools/kinoath

[13] http://class.uni-koeln.de

[14] http://tla.mpi.nl/tools/tla-tools/yams/

used to, Arbil is overall a great help in making the task of metadata creation easier and faster. It also assists in making better quality metadata. The use of Controlled Vocabularies, fixed formats for data entry, and the ability to add repeated information from the Favourites directory all help to write more consistent metadata. The structure it provides for session nodes also encourages very detailed metadata, which can easily be added to at any time. If the archive you are using requires IMDI or CMDI format metadata, Arbil is the program to use to create it. Even if you do not require IMDI or CMDI format metadata, I would recommend Arbil since the drawbacks of unnecessarily meeting the IMDI/CMDI requirements are far outweighed by the benefits Arbil provides, such as the ability to save elements to the Favourites directory, the automatic reading of media and written file information, and the knowledge that you can edit your metadata easily if needed later. If you need to create metadata in another format, Arbil may not be the best choice, though it should be possible to create a custom template to suit this format. Overall, I think Arbil is an excellent tool for creating, editing, and searching metadata, and I look forward to its future improvement.

Primary function:	Creating, editing and searching metadata in the IMDI or CMDI format.
Pros:	Assistance in entering information via drop down lists, the Favourites directory, automatic entering of resource file information, and the ability to edit multiple fields at once. The Find And Replace function, which assists with editing data. The ability to open files directly from Arbil.
Cons:	Steep learning curve. It is designed for use with IMDI and CMDI metadata formats only, though it is possible to create new templates in Arbil to enable use of other XML formats.
Platforms:	Web-based, PC/Windows, Mac OS X, Debian Linux.
Open source:	Yes, the source code is available upon request under the GNU General Public License, Version 2.
Proprietary:	Users can use Arbil under the GNU-GPL license.
Available from:	The Language Archive, Max Planck Institute for Psycholinguistics (http://tla.mpi.nl/tools/tla-tools/arbil/).
Cost:	None
Reviewed version:	2.4.36550

Application size:	Windows XP or higher: 36 MB; Mac OS X: 22 MB; Ubuntu 12.10: 20 MB.
Documentation:	http://tla.mpi.nl/tools/tla-tools/arbil/

References

Bowern, Claire. 2008. Linguistic fieldwork: a practical guide. Basingstoke: Palgrave MacMillan.

Dima, Emanuel, Erhard Hinrichs, Christina Hoppermann, Thorsten Trippel & Claus Zinn. 2012. A metadata editor to support the description of linguistic resources. In Calzolari, Nicoletta Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation at LREC 2012*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/468_Paper.pdf

Saad, George, Anastosios Vogiatzis, Katarzyna Wojtylak, Claudia Zabeo. 2012. Arbil for editing and managing IMDI metadata Version 2.4. http://www.mpi.nl/corpus/html/arbil-imdi/index.html. Last updated 20-12-2012.

Thieberger, Nicholas & Andrea L. Berez. 2012. Linguistic data management. In Nicholas Thieberger (ed.) *The Oxford handbook of linguistic fieldwork*, 90-120. New York: Oxford University Press.

Withers, Peter. 2012. Metadata Management with Arbil. In Arranz, V. D. Broeder, B. Gaiffe, M. Gavrilidou, & M. Monachini (eds.), *Proceedings of the Workshop Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR at LREC 2012*. 72-75. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/workshops/11.LREC2012%20Metadata%20Proceedings.pdf#page=79

Rebecca Defina
rebecca.defina@mpi.nl