IMDI Metadata Field Usage at MPI

Alex Klassmann, Freddy Ofenga,
Daan Broeder, Roman Skiba
MPI, N megen

Introduction
Metadata is indispensable for discovering and searching the ever-growing volume of online language resources. Three metadata standards are now widely used for language resources - TEI, OLAC, and IMDI (links below). TEI is the oldest of these; OLAC was developed as an extension of the Dublin Core (DC) set which is widely used by librarians and for generalised cataloguing of web documents. The IMDI set was designed in collaboration with linguists, speech engineers and others to serve the specific needs of those researchers, especially resource discovery and retrieval, and is correspondingly more comprehensive.

An implicit purpose, therefore, of using IMDI is to support more accurate retrieval of resources. In reality, however, this can only be achieved if the metadata fields are actually accurately populated with searchable content. A large number of empty or inappropriately filled-in fields would prevent enhanced retrieval. Therefore, we saw that, after six years in operation, it would be very interesting to analyse our depositors’ usage of IMDI. From such a study, we felt we could better understand:

• how well searches are working (searches that depend on poorly used elements may lead to wrong interpretations);
• where researchers find it difficult to enter descriptions and where, therefore, improvements to IMDI could be made;
• why some researchers complain about the necessity to create metadata (for example, some PhD students complain that time pressures do not allow them sufficient time to do so).

We focused on metadata descriptions that were created by individuals or small projects from the MPI and from DoBeS teams. Corpora where metadata was completely or partly generated, such as the Dutch Spoken Corpus, were excluded from the study. A total of 23,710 metadata description files were analysed.
Results
Figure 1 gives an overview of depositors’ usage of IMDI fields (where usage means that some data has been filled in). A number of observations can be made:

At the session level, project, geographic and date information is filled in for about 90% of cases, but descriptions with further useful information are provided in only 40% of cases.

The description field at the content level is used in more than 70% of cases. At this level, depositors prefer to use this free-text description field rather than the Content Type fields such as Genre (30%), Sub-Genre (25%) or Subject (10%). The modalities in focus and the communicative context are used in more than 75% of cases.

The language name is filled in in almost 100% of the cases. However, the language code was used in only 40% of cases, even though many of the codes can be selected from supplied lists.

Information is frequently provided about actors. On average there are three actors (including the creator) per resource bundle. Language skills of informants are filled in in many cases, but information about sex, age etc. is very limited.

Information about references, formats and types is available for almost 100% of resources. This means that the IMDI records do indeed act as a kind of glue bundling together files. In addition, we can use such information to automatically check consistency, e.g. for correct file extensions. Some fields such as file size are little used but could be filled in automatically.

Discussion
The poor usage of the content type fields is somewhat disappointing. Local discussions have revealed that depositors find it difficult to use the built-in vocabularies and value sets, and that they have problems with categorising their resources. Some depositors did not know how to use these fields, or that it is possible to select multiple values. We doubt whether ad hoc changes in the value sets of the Genre, Subgenre or Subject fields will improve the situation, since we have to conclude that there is no commonly accepted vocabulary for them (except for some very basic terms). At a recent DoBeS workshop (June 2006), some researchers argued that classifications in endangered languages documentation would be more useful if they included genre vocabularies as understood by the language communities themselves.

The statistics also made us look in more detail at the use of the language name and language code fields. It was not clear to us why there was such a large discrepancy between the rate of usage of language names and language codes. Further investigation showed that in most cases it was possible to select a suitable language code. In fact, we developed scripts to correct obvious mistakes and add missing entries in these fields. As a result, the language name and code fields are now filled in and consistent in almost 100% of cases. We assume that depositors are either unfamiliar with the Ethnologue language codes or that they do not feel comfortable using them.

We concluded that the most important IMDI fields such as location, language, and recording date are used at a satisfactory level across all resources. For other fields, it seems that usage is dependent on the type of collection, and a more elaborate analysis needs to be carried out. However, the study has made clear that the description of the content type (Genre, Subject, etc.) can’t be done at a satisfactory level. We were not yet able to draw any conclusions about particular IMDI fields that might reasonably be eliminated.

![Figure 1. Usage of the most relevant IMDI fields, showing the proportion of IMDI files for which the field was filled in by depositors (from Klassman et al., 2006).](image-url)
References and links


LEI: http://www.tei-c.org/
OLAC: http://www.language-archives.org/OLAC/metadata.html
Dublin Core http://dublincore.org/
IMDI: http://www.mpi.nl/IMDI