# 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 specifc 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 felds are actually accurately populated with searchable content. A large number of empty or inappropriately flled-in felds 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 fnd it difcult 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 sufcient 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 fles were analysed.

**Results**

Figure 1 gives an overview of depositors' usage of IMDI felds (where usage means that some data has been flled in). A number of observations can be made:

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

The description feld at the content level is used in more than 70% of cases. At this level, depositors prefer to use this free-text description feld rather than the Content Type felds 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 flled 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 flled 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 fles. In addition, we can use such information to automatically check consistency, e.g. for correct fle extensions. Some felds such as fle size are little used but could be flled in automatically.

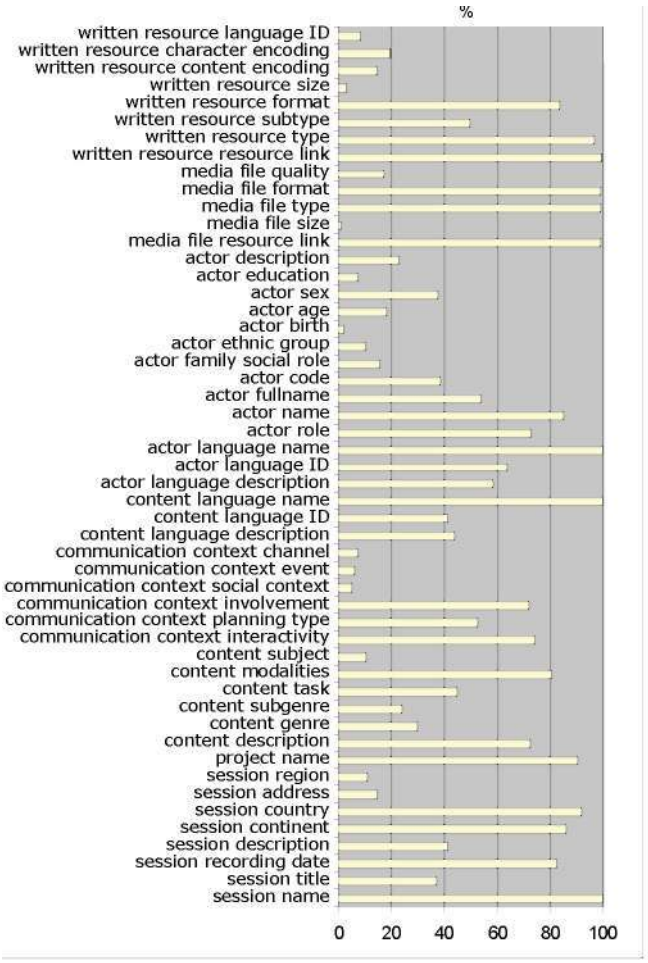


*Figure 1. Usage of the most relevant IMDI felds, showing the proportion of IMDI fles for which the feld was flled in by depositors (from Klassman et al., 2006).*

**Discussion**

The poor usage of the content type felds is somewhat disappointing. Local discussions have revealed that depositors fnd it difcult 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 felds, 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 felds 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 classifcations 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 felds. 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 felds. As a result, the language name and code felds are now flled 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 felds such as location, language, and recording date are used at a satisfactory level across all resources. For other felds, 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 felds that might reasonably be eliminated.

## References and links

Klassmann, A., Ofenga, F., Broeder, D., Skiba, R., Wittenburg P. (2006). Comparison of Resource Discovery Methods. In: N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik & D. Tapias (Eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 113-116. Paris: European Language Resource Association.

Wittenburg, P., Peters, W., Broeder, D. (2002). Metadata Proposals for Corpora and Lexica. In: M. G. Rodriguez & C. P. S. Araujo (Eds.), *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 1321-1326. Paris: European Language Resource Association.

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