Comparison of Resource Discovery Methods

Alex Klassmann, Freddy Offenga, Daan Broeder, Romuald Skiba, Peter Wittenburg

Max-Planck-Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen
{alex.klassmann,freddy.offenga,daan.broeder,romuald.skiba,peter.wittenburg}@mpi.nl

Abstract

It is an ongoing debate whether categorical systems created by some experts are an appropriate way to help users finding useful resources in the internet. However for the much more restricted domain of language documentation such a category system might still prove reasonable if not indispensable. This article gives an overview over the particular IMDI category set and presents a rough evaluation of it's practical use at the Max-Planck-Institute Nijmegen.

1. Introduction

The raw material for linguists are samples of a particular language. These may range from pieces of parchment til recordings of TV broadcast. Although there exist guidelines for the metadata description and annotation of linguistic resources (IMDI [1], DC/OLAC [2], TEI [3], EAGLES [4], specialized data bases), no standard is universally accepted and probably can't be since researchers will focus on different aspects and invent new theories and ideas. The amount of collected and electronically available resources has exploded over recent years and poses the problem of organization/management and (re-)discovery of the data. In this paper we will present the approach the MPI for Psycholinguistics has chosen with respect to the metadata description, will elaborate on a number of different location methods and finally will discuss some critical points. The first paragraph will give a short overview over the IMDI metadata scheme. Than their practical application i.e. the tools which allow the user to handle this metadata set will be presented. A rough evaluation of the quality of the at present available metadata follows. Than an alternative to formal categorization will be presented, namely free „tagging“, which is currently lively discussed with respect to internet search engines. It's applicability to the field of linguistics will be questioned and some preliminary conclusions drawn.

2. IMDI Metadata

The IMDI (ISLE MetaData Initiative) scheme was developed during 2001-2003 by a broad network of linguists from different sub-disciplines such as field linguistics, phonetics, multimodality research and corpus linguistics. Its purpose is to give a solid, precise and extensible framework for the organization, bundling and retrieval of in principle any kind of digital linguistic resources, in particular annotated media streams and text sequences making up by far the largest percentage of current resources in language resource archives.

Typically primary language documents like audio or video files are accompanied by one or more text files, containing a transcription, translations and annotations at other linguistic levels (morphosyntax, semantic, etc) of the former and seen in the IMDI framework as resources themselves. An IMDI-session contains a detailed meta description of those tightly connected resources, and could therefore be named equivalently as metadata about a 'resource bundle'. The IMDI-schema describes in addition how those sessions can be grouped together into corpora and sub-corpora. Although corpus organization is relevant for management and browsing, it is not of relevance in this paper, i.e., for more details we refer to other IMDI documentation [5,6].

An IMDI-session can be best thought of as a form with roughly 150 hierarchically ordered entries, which concern e.g. information about

- the event (recording location, date, etc),
- the languages involved,
- the speaker(s),
- the type and nature of speech,
- technical information about the resources and
- access rights.

For most fields one or more values can be selected, but there are also so-called descriptive fields for the input of free text. Furthermore there is the possibility for every user to add arbitrary key-value-pairs which can be interpreted as a personal or project-specific extension of the schema. In order to facilitate the procedure of filling in the metadata, a special professional editor has been build at the Institute.

A single field, the „bundle name“-field is obligatory, yet users are urged to fill in all others, too. Unfortunately they tend to avoid this time-consuming work oriented to a re-usage by others and fields stay empty or have a default setting. Although everyone agrees that filling in metadata is very important in many respects, in particular since the knowledge about the content may be lost within shortest time, the amount of time spent on this aspect in the whole resource management life cycle is still too little.

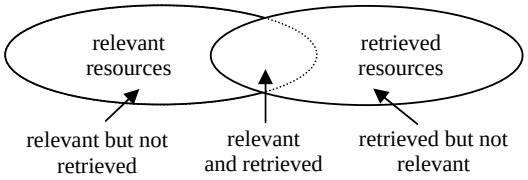
3. Methodological Issues

One important question for the usage of archives – traditional as well as modern – with an extremely growing amount of resources is the possibility for the user to locate useful resources. As described the MPI uses the structured IMDI set to describe resources which therefore lends itself to carry out queries. Metadata includes added value with respect to the resources themselves, therefore it is data that cannot be missed. A recording may include an interview with a person having certain characteristics such as age, sex, education etc. Only in rare situations the

recording will contain this information explicitly – it is the metadata description that will allow the interested user to make a comparison between male and female language use for example. Many other examples of this added value can be given.

Although we will have very different user groups ranging from researchers, teachers, students, journalists to the speakers themselves. All have different types of queries and all asking different types of interfaces. Nevertheless, we can make a few general statements on what a typical search method should optimize.

Literature defines two terms, “precision” and “recall”, as measures for the success of a query. With “precision” the proportion of hits that are relevant compared to the irrelevant hits is meant. A higher amount of “noisy” results would therefore reduce the precision rate. With “recall” the proportion of relevant hits that were found compared to the not found relevant hits is meant. A query method that would not find very much of the relevant resources a user is looking for obviously would be not successful. The following drawing taken from G. Simons [7] is very useful to indicate the relation between the two terms.



Another important point in searching is of course the question of how to rank the hits. The precision could be very low, i.e., the number of irrelevant hits could be high, but if the relevant resources would be presented at the top of the list the user probably wouldn't bother. In this paper we will not discuss the ranking aspect.

4. The MPI Archive

The Max-Planck-Institute Nijmegen houses a digital archive with a large variety of different language corpora, all categorized with the IMDI metadata set. The archive encompasses ca. individual 45.000 IMDI-sessions describing about 150.000 resources.

Infrastructure and tools have been designed to offer to the user several options to search for a specific IMDI-described resource. Since metadata is open per definition, all descriptions are accessible via the web; cf. http://corpus1.mpi.nl/ds/imdi_browser):

1) Browsing in linked resources. This is similar to clicking through a local file system with the difference that the hierarchy of corpus structures is much more stable. The approach is aimed at users familiar with or quickly able to grasp the underlying logical organization. Bookmarks help to make this process more efficient.

2) Structured search within the whole archive as well as within a selected part of it. Every IMDI-element can be addressed individually and the search for different elements can be combined into one query. Queries like "Give me all video files that show a female Wichita speaker older then 60 years" can be formulated and a high precision, i.e., a low number of irrelevant hits, can be expected. Yet, the user has to know the terminology used by the IMDI schema in order to achieve a high recall, i.e., get a high percentage of the resources having looked for as hits. Furthermore, search is restricted to elements with closed or open vocabularies and does not cover elements with free text.

3) Unstructured search over the whole or part of the archive. The user can enter words or regular expressions into a free text field (Google-like). Any metadata element including the free text descriptions that contains matching strings will produce a hit. It is possible to formulate logical combinations of expressions and even "fuzzy terms" (for an overview of the possibilities cf. [8]). The recall with this method can be expected to be higher compared with structured search, however, the precision will be poor, i.e., much more irrelevant hits can be expected.

4) An extension of unstructured search is to provide the metadata descriptions to web search engines like Google with their advanced information retrieval techniques. However search cannot be restrained to a specific corpus, not to mention parts of it, and results will include a huge amount of unwanted hits from the whole internet. An additional term such as „IMDI“ or „MPI“ improves the precision significantly, but still yields unsatisfactory results.[1]

5) All IMDI records were transmitted to the OLAC service provider (DC [9]). OLAC offers a structured search possibility, but limits itself to the elements of DC and a few additional ones such as the language a resource is in. Currently, the service is not working well, since the OLAC service provider does not accept too many records, i.e., they expect the data provider to just deliver one metadata record for a sub-corpus. For the MPI it is in many cases difficult to determine what exactly a sub-corpus is. With respect to precision and recall we expect similar results as with structured search, as long as the restricted set of elements is sufficient. An advantage of using OLAC, however, is that other archives will contribute to OLAC, too.

6) Geographically orientated browsing. Since many languages in the archive are related to diverse and less known regions all over the world, a geographical browsing makes sense, too. The visualization tool Google-Earth [10] is used for this purpose, where the user can look for spots on the physical map of the Earth that point to IMDI-files. Of course, this method yields an enormous high precision and recall if only the geographic location is the discovery criterion. Since this approach is of less theoretical interest, we will not elaborate on this option.

[1] When searching for example for real resources for the TEOP language a Google search with “teop” as query string yields 17.600 hits with lots of unusable hits. A query string “imdi teop” only yields 683 hits and more important the entry for the Teop corpus is amongst the first five. However, users suffer from the same deficit: how should they know which string to use to achieve an acceptable precision and recall.

We should not forget to mention that in general researchers want to combine metadata search/browsing with searching on the content as it is possible now for example with ANNEX [11]. Typical questions such as "give me all instances where a 4 year old female speaker is using a certain morpho-syntactic construction" can only be addressed when a combined structured search is performed. But we also understand that such questions will only be addressed by the "very well informed" user who knows exactly the terminology that is used. All other search options will not lead to useful results. In this paper we will not include the content search option, but discuss metadata search options in general.

5. Evaluation

In order to have significant variance in the data, an evaluation of the metadata was done on a subset of the resources in the archive, where metadata was filled in manually and by different users, i.e., the Dutch Spoken Corpus, for example, was not included.

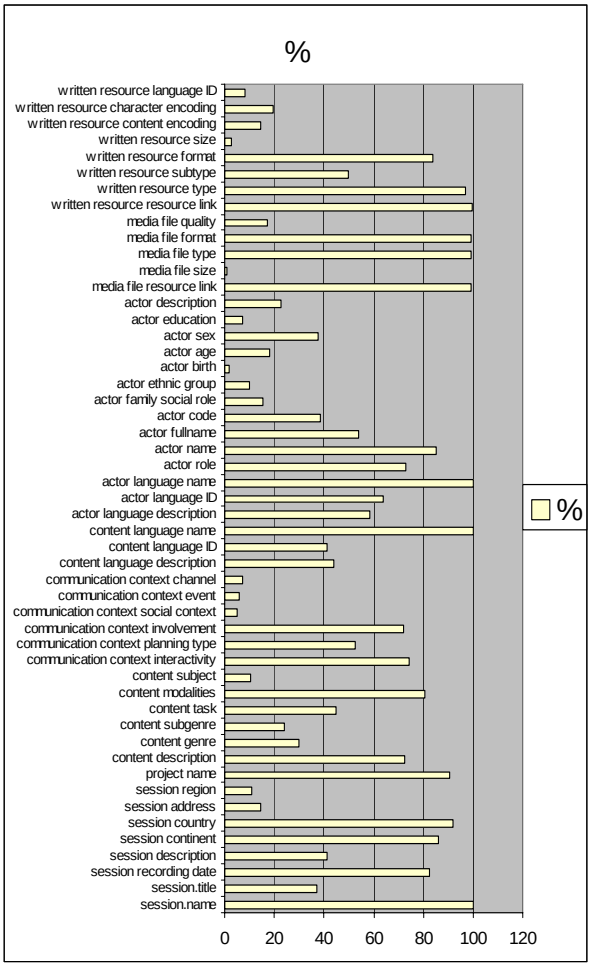
The table below gives an impression of how often fields are actually filled in (e.g. not empty and not default values like „unknown" or „unspecified"). These statistics were created on 23.710 resource bundles. As can be seen the sets are far from being complete. On the other hand, every field of the scheme (including those not shown in the table) has been used in some sessions, so that it seems that no field in the schema is obsolete. These statistics give a baseline idea of what can be expected.

Since there is still not sufficient experience at the institute with actually performed metadata searches, it is not yet possible to carry out a full-fledged statistical evaluation based on empirical data. Instead, test queries which might be of relevance for researchers were formulated and executed. It was then checked whether the hits were accurate.

So, e.g. in Second Language Learning Research the influence of age on the acquisition of language is examined and it is assumed that there is a critical period in childhood for the development of certain skills such as learning grammar constructions. In order to find resources one would like to formulate a query like „Give me all resources for a given (not-mother-)language for speakers aged between 4 and 16 years". Since the development between boys and girls may differ one even could refine the query by an appropriate qualifier.

Using the IMDI structured search the following query "Language=Dutch, Actor.Language.Mothertongue=false, Actor-Age<16 and >4" yields 203 hits. An additional selection on "Actor-sex = Male" results in 119 hits and one with "Actor-sex = Female" in 83 hits. A full-text search with a query "Dutch AND second AND language AND (15 OR ... OR 5)" results in 488 hits and may be still useful, too.

Categorization with respect to age and sex as well as technical categories like the file format are rather uncontested and not prone to subjective interpretation.



This is different with respect to the descriptive elements concerning the content. Here the difficulty can be seen at the many corrections the initial IMDI set experienced and the user is merely offered a list of given values, but can type in others ("open vocabulary").

The vocabulary for the element „Content-Genre" e.g. encompasses 13 items („discourse", „poetry" etc.), two of them never have been used („Popular fiction", „Newspaper article") and another 15 values have been added by users. Concerning the element „Content-SubGenre" the situation is similar: no offered type of drama has been used and (fortunately!) no resource was classified as „Unintelligible Speech". Some 30 items were added, ranging from broad terms like „Speech" to very specific ones. This poses the question if such a categorization in advance by a group of „experts" is the right approach for data organization.

6. Free Tagging

In this paragraph we will discuss free user „tagging" as opposed to categorization based on an a priory defined categorization schemes.

With respect to searches in the internet the early stage approach from Yahoo to perform search along given categories has been abandoned in favour of key word

search as known by Google. Yet simple string matching in documents is not very precise and doesn't work at all for media files. Currently an alternative to in-advance categorization might be 'user tagging' as it is promoted most outstandingly be Shirky [12]. He refers to a service [13] that offers users to store bookmarks of web-resources and make those bookmarks available for the public. So each user who wants to remember an URL of interest can describe it with an arbitrary set of key words. Of course, each user has his own view of the resource and the description may be inaccurate or erroneous, but the assumption is, that if there are a lot of users describing the same URL, the statistics will end up establishing a widely shared set of key terms. This kind of „categorization afterwards" lacks genuinely any hierarchy and results more in a kind of semantic net or „topic map".

7. Discussion

There are a number of reasons why the idea of "free tagging" will not be applicable for the domain of language resources:

- The idea of „free tagging" relies on the voluntary work of many and presupposes that the resource in question is interpretable by everybody. This is certainly not the case in the field of linguistic data, where often only the producer of the resource is able to describe it adequately.
- It is the researcher who has the deep knowledge about the construction of a corpus and about the reasons to have chosen a certain approach. This knowledge has to be stored somewhere and it's the metadata where it is stored.
- At least the linguistic users can rely on the a priori defined categorizations, since linguistic terminology has stabilized to a large extent during the last decades.

So, tagging of the content of linguistics resources would have primarily to be done by the creator like with the rest of the metadata. On one side, the „open vocabularies" offered currently by IMDI incite some users to slightly misuse them for an imitation of „free tagging" e.g. if they add an overspecialized item. On the other hand "free tagging" could be an option for other "experts" to enrich the data and therefore to increase the precision and recall.

A solution and kind of promise between the two strategies may be to make every new entry „public", e.g. adding it to the list of offered vocabulary automatically. This would benefit those who fill in the data as well as those who are querying it. Furthermore, it would inhibit users to add too specific terms by a kind of „social pressure".

8. Conclusion

The Max-Planck-Institute Nijmegen offers several kinds of querying and browsing approaches corresponding to different user interests. The IMDI categorization scheme allows in principle for very detailed search and therefore has the potential for a high precision and high recall compared to all sorts of free text searches.

However, the IMDI forms are generally not completely filled in as was indicated in the table and even linguistic users do not fully share the same terminology. This will deteriorate the success of the searches in terms of precision and recall. Since free-text field also bear relevant information in many cases, even some linguists will prefer nevertheless a free-text search on the metadata first.

9. References

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

[2] http://www.language-archives.org/

[3] http://www.tei-c.org/

[4] http://www.ilc.cnr.it/EAGLES96/

[5] Wittenburg, P., Peters, W., Broeder, D. (2002). *Metadata Proposals for Corpora and Lexica.* In M. Roriguez Ganzalez & C. Paz Suarez Araujo (eds.), Proceedings of the 3rd International Conference on Language Resources and Evaluation. Paris: European Language Resource Association. pp 1321-1326

[6] Broeder, D., Wittenburg, P., Crasborn, O. (2004). *Using Profiles for IMDI metadata creation.* In X. Fatima Ferreira et. al. (Eds), Proceedings of the 3rd International Conference on Language Resources and Evaluation. Paris: European Language Resource Association. pp1317-1320

[7] Aristar-Dry, H., Simons, G. (2006). *E-Meld: openness, ontologies and interoperability*. DGFS Annual Meeting on Language Documentation and Description – Working Group 6. University of Bielefeld

[8] http://lucene.apache.org/java/docs/queryparsersyntax.html

[9] http://dublincore.org/

[10] http://earth.google.com/

[11] http://www.mpi.nl/annex

[12] Shirky, Clay (2005): www.shirky.com/writings/ontology_overrated.html

[13] http://del.icio.us