Originally published in: Hedeland, Hanna/Schmidt, Thomas/Wörner, Kai (Eds.): Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011.
– Hamburg: Universität Hamburg - SFB 538, 2011. Pp. 105-109. (Arbeiten zur Mehrsprachigkeit : Folge B 96)

Devil's Advocate on Metadata in Science

Christina Hoppermann, Thorsten Trippel, Claus Zinn
General and Computational Linguistics, University of Tübingen
Wilhelmstraße 19, D-72074 Tübingen
E-mail: christina.hoppermann@uni-tuebingen.de, thorsten.trippel@uni-tuebingen.de, claus.zinn@uni-tuebingen.de

Abstract

This paper uses a devil's advocate position to highlight the benefits of metadata creation for linguistic resources. It provides an overview of the required metadata infrastructure and shows that this infrastructure is in the meantime developed by various projects and hence can be deployed by those working with linguistic resources and archiving. Possible caveats of metadata creation are mentioned starting with user requirements and backgrounds, contribution to academic merits of researchers and standardisation. These are answered with existing technologies and procedures, referring to the Component Metadata Infrastructure (CMDI). CMDI provides an infrastructure and methods for adapting metadata to the requirements of specific classes of resources, using central registries for data categories, and metadata schemas. These registries allow for the definition of metadata schemas per resource type while reusing groups of data categories also used by other schemas. In summary, rules of best practice for the creation of metadata are given.

Keywords: metadata, Component Metadata Infrastructure (CMDI), infrastructure, sustainable archives

1. Introduction

The creation of primary research data and its analysis is a large share of a researcher's workload. In linguistics, research data comprises many different types: there are resources such as corpora, lexicons, and grammars; there are various kinds of experimental data resulting, for example, from perception and production studies with sensor data originating from eye-tracking and MRI (magnetic resonance imaging) devices. There is data in the form of speech recordings, written text, videotaped gestures, which, in part, is annotated or transcribed along many different layers; there is audio and video data of other forms of human-human communication such as cultural or religious songs or dances; and there is also a large variety of software tools for the manipulation, analysis and interpretation of all these types of data sources.

Once a study of research data yields statistically and scientifically significant results, it is documented and published, usually complementing a description of research methodology, interpretations of results, etc., with a depiction of the underlying research data. Reputable journals are archived so that its articles are deemed accessible for a long time. Access to articles is usually facilitated via Dublin Core (DC) metadata categories such as "author", "title", "journal", "publisher" or "publication year". In general, however, there is no infrastructure in place to access the research data underlying a reported study, although some researchers make such data available via their webpage or institution, and some conferences or journals ask authors to supplement their article with primary data, which is then also made public.[1] So far, it is not the general rule to describe research data with metadata for indexing or cataloguing by themselves or others. In part, this is due to caveats for the provision of metadata held by large parts of the scientific community. In this paper, the Devil's Advocate (DA) will articulate some of these caveats. We will aim at rebutting each of them, given the recent advances for metadata management, in particular, in the area of linguistics.

2. Playing Devil's Advocate

DA: There is little if any scientific merit to be gained from resource and metadata provision.

This is a view mentioned in a recent statement by the Wissenschaftsrat[2] which says that infrastructure does

[1] For example, Interspeech 2011 invited authors to submit supporting data files to be included on the Proceedings CD-ROM in case of paper acceptance.
[2] The German Wissenschaftsrat is a joined council of German

hardly provide for an increased scholarly reputation (Wissenschaftsrat, 2011:23). Though this might be true for a restricted notion of scientific merit, that is the merit being defined by the number of published journal articles and books, it is not true in a less restricted sense. Furthermore, the Wissenschaftsrat (Wissenschaftsrat, 2011:23) points out that infrastructural projects offer the opportunity for methodical innovations, generate new research questions, and help attracting new researchers. If new researchers, methods and research questions are part of the scientific merit, the claim that there is no scientific merit in metadata provision is thus not true. There are even more reasons for arguing that additional scientific merits are gained, at least in three overlapping areas: (1) by providing a complete overview over the field, (2) by fostering interoperability and providing reproducible, non-arbitrary results, and (3) by increasing the pace of gaining research results.

First of all, in an ideal case, a metadata-driven resource inventory gives an accurate picture of a scientific landscape by containing all resource types such as corpora, lexical databases, or experiments. By having access to all these resources, in principle, nothing is gained because it is too time-consuming to analyse and reproduce research questions from the data. But as soon as resources are described by metadata, it is possible to classify, sort and provide an overview over them using the descriptions as such. Though descriptions contain generalisations, they are still sufficient to provide an outline of resources. This also serves the purpose of providing essential background for steering research activities and funding projects as well as to discover trends and gaps, all allowing to increase the researcher's reputation and merit.

Second, the metadata-based publicity fosters communication between researchers, for example, because contact information are required to gain access to resources, comparable data structures are needed to be reusable by other methods, or because selections of resources (e.g. subcorpora) have to be created. Resources can be merged and cross-evaluated to discover which results are reproducible. This helps to avoid fraud and plagiarism. At the same time, the investigation of research questions different from the original ones can be applied to existing resources. In all cases, good scientific practice will credit the resource creator, and thus add to his or her reputation when a publication makes reference to its underlying research data, which is possible on the basis of appropriate metadata. The references pointing to the resources can be indexed by others and are consequently added to the scientific map.

Third, more and faster results can be created. By providing metadata, researchers new to a discipline gain a faster overview over the research questions and activities of a discipline as well as easier access to existing linguistic resources and tools. Moreover, accurate metadata descriptions can help avoiding the duplication of research work by providing insights and access to existing work. Hence, researchers who are applying new methods do not always have to recreate resources but can rely on existing ones, providing a jumpstart. At the same time, the resources as such are providing added benefit by being more widely used, thereby also increasing the reputation of the creator.

DA: Expert knowledge on metadata is required to properly describe research data. Thus metadata experts rather than researchers are called for duty.

The library sciences, with their long tradition and expertise in metadata, have many different classification systems in place to organise collections. But is it realistic to ask researchers, such as linguists, to properly describe language resources and tools with metadata, given their lack of knowledge in metadata provision, the variety and complexity of research data, and the missing dominant metadata schemes in the field? On the other hand, it seems clear that metadata provision cannot be done properly without the researchers' involvement. It is unrealistic to assume that some research data can be just given to a librarian with expertise in linguistics (or a linguist with expertise in archiving methodology) with the task to assign proper metadata to it. There needs to be considerable involvement of the resource creator in describing the resource in formal (where possible) and informal terms (possibly by filling out a questionnaire). The "librarian" can then enter the provided information into a formal schema, ensuring that, at least, obligatory descriptors are properly provided. In sum, to put a proper metadata-based infrastructure in place, some minimal researcher training in metadata provision is needed. This

Research Foundation officials and researchers appointed by the government for consulting it on research related issues.

needs to be complemented with infrastructure personnel, or, if possible, with user-friendly metadata editors that trained researchers can learn to use.

DA: There is a little if any consensus on the set of metadata descriptors or metadata schemes to be used in describing language resources and tools.

It is clear that a common vocabulary for metadata provision is required. Otherwise it will be hard to offer effective metadata-based search and retrieval services. It is also evident that established metadata standards such as Dublin Core are insufficient, as they do not include every data category (DatCat) needed for describing specific types of resources. However, given the complexity of the research field in linguistics with its many different resource types, it is naïve to assume that established metadata schemas can be reused without loosing descriptive power. For example, resource types need to be indicated and for different resource types additional descriptive categories need to be defined. For lexical resources it is common to describe the lexical structures, for annotations the annotation tag sets, for experiments the size of the samples and the free and bound variables. Each of these data categories is only relevant for the individual type of a resource, but for these they can be more essential than categories such as "title" and "author". As this list of data categories may require additions, since new resource types become available, it needs to be treated as an open list.

In recent times, some consensus on the procedure of creating elementary field names for the description of linguistic research data has been achieved in order to allow for a standardisation of data categories. It is formally captured by the ISOcat data category registry for the description of language resources and tools (ISO 12620; International Organization of Standardization, 2009; http://www.isocat.org). ISOcat (Figure 1) is an open web-based registry of data categories into which everybody can insert his own data categories with (human-readable) definitions of their intended use. This is done in a private space with limited access that can be used by researchers to include new data categories not yet intended or not ready for standardisation. For private use, these data categories can already be referenced via persistent identifiers (PIDs) but they can also be stored in a public space with unrestricted access and be proposed as standard data categories. If the data categories are submitted for standardisation, a standardisation process involving domain experts is being initiated with community consensus building, quality assurance, voting and maintenance cycles.

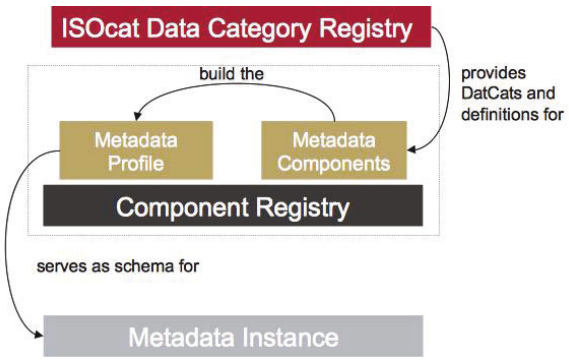

Figure 1: Relation between ISOcat, Component Registry and metadata instances

The registry provides a solid base to start from, but the sheer size of available DatCats may overwhelm untrained users. Additional structures are needed to minimise cases where different users may apply different descriptors to provide similar resources with metadata. For this purpose, the Component Registry for metadata (Figure 1; Broeder et al., 2010; http://catalog.clarin.eu/ds/ComponentRegistry/#) contains a rich set of prefabricated metadata building blocks that aggregate elementary blocks of data categories into larger compounds. Researchers can select and combine existing buildings blocks – or define new ones – in a schema, which can then be instantiated to describe a given resource with the help of a so-called metadata instance (Figure 1). The concept of reusing building blocks is part of the Component Metadata Infrastructure (CMDI, http://www.clarin.eu/cmdi). For many resource types the registry already contains prefabricated schemas that can be re-used by researchers. Moreover, there exists at least one fully functional metadata editor (http://www.lat-mpi.eu/tools/arbil/) with interfaces to both ISOcat and the Component Registry. It is freely available and support is provided by the programmers to facilitate the use of the editor for non-expert users who otherwise might be overwhelmed by the total range of functions the editor offers. There are also other XML editors supporting the schemas. Once a schema is defined with these tools, these off-the-shelf

XML editors are available to describe resources with metadata according to the metadata schema. These schemas can then be used to validate the metadata instances with the help of syntactical parsers to ensure the adherence to syntactic structures and controlled vocabulary.

DA: There is rarely a right time to make a resource public (via metadata description).

Research rarely follows a fully planned path. A resource such as a corpus or a lexicon is adjusted, additional layers of annotation or transcription are added, data may get re-annotated with different coders, lexical entries may get revised or extended to reflect new insights, etc. Nevertheless, the moment publications are created and project reports are written, it shall be good scientific practice to archive the underlying research data and to assign and publish metadata about the resource. Here, the current status of the resource can be marked with metadata about, for instance, the resource's life cycle or versioning information.

There is also a policy change in the funding agencies. The German Research Association (DFG), for instance, sets the terms that resources ought to be maintained by the originating institution; researchers are responsible for the proper documentation of resources, and procedures need to be defined for the case when they leave an institution (Deutsche Forschungsgemeinschaft, 1998:13). A proper documentation of resources has to include their description in terms of metadata to facilitate their archival and future retrieval.

Therefore, at the latest, metadata shall be provided (or revised) at the end of a research project, at best by the researchers who have created the resource. Ideally, the life cycle stage at archiving time is already defined in the project work plan. Even if the desired final state was not accomplished, the primary data needs to be archived by the end of the project with proper metadata assigned to it.

DA: Without a central metadata agency, all the added values advertised will not materialise.

Added values such as searchability and citation of resources require some point of access to the metadata. It is correct that there is not a single central metadata agency but there are various interconnected agencies providing services to the community in terms of metadata. For instance, the German NaLiDa project (http://www.sfs.uni-tuebingen.de/nalida/) serves as a metadata centre for resources and tools created in Germany. The project as such does not claim exclusive representation, but aims at cooperating with other archives in providing a service to the community for accessing metadata in the form of catalogues and allowing easy access to resources. It harvests metadata from participating institutions and also provides metadata management support for German research institutions (Barkey et al., 2011). Within the project, a faceted search interface was developed with complementation of a full-text search engine (http://www.sfs.uni-tuebingen.de/nalida/katalog), with currently access to more than 10,000 metadata records of language resources and tools. Though the NaLiDa project could be seen as a central metadata agency, its implementation has a rather decentralised flavour. Metadata is harvested from various sources and then aggregated and indexed into a single database. To kick-start or increase the inflow of data, participating institutions receive help both in terms of setting-up an OAI-PMH[3]-based data provision service and in other aspects of metadata creation and maintenance. Once the local metadata providers – the primary research data remains with the institutions – are set up, other parties than NaLiDa are free to crawl their data sets and to provide services in terms of all data.

At the European level, the CLARIN project (http://www.clarin.eu) has also devised such a crawler, and is likewise offering a faceted search interface for language resources and tools (CLARIN Virtual Language Observatory, http://www.clarin.eu/vlo/). Since both (and other) parties work towards the realisation of a common infrastructure, with different foci but similar goals, there is much to be gained from a healthy competition and exchange of ideas for the scientific community to profit from.

3. Summary

Given the recent advances in linguistics with regard to metadata provision for linguistic resources and tools, there is little left to offer excuses for not using the existing infrastructure. In general, this results in the

[3] Open Archives Initiative Protocol for Metadata Harvesting

following rules of best practice for the documentation of resources:

1) One of the best strategies for preserving research data is by publishing it into repositories and networks. This way, multiple archives serve as backup. Additionally, it allows for an easier sharing and spreading of resources, contributing to the academic merits of resource providers.
2) Archived data is easier accessible if the data is sufficiently described. As flexible metadata schemas can adapt for various types of resources, it is possible to create such descriptions as required by the type of a resource. Metadata can then be used to make resources public, in order for others to use (harvest) them.
3) Data categories are best defined in central (standardised) registries, such as ISOcat, that allow for references via persistent identifiers. No data categories should be used that are not centrally defined to avoid fragmentation of the resource community.
4) For interoperability purposes, components as collections of data categories should be reused where adequate or defined as new entries in the Component Registry for reuse by others.
5) The flexibility of the framework helps to avoid tag abuse if data providers adhere to data category definitions or, if not available, define their own modified categories. This will contribute to the consistency and reusability of data.
6) Syntactic evaluation of metadata should always be performed to ensure harvesting, usability of applications and consistency. By checking for content models, tag abuse can be avoided further.
7) When using research data, it should be referred to them as stated in the data's metadata.
8) Resource creators might need some training and assistance, which is provided by various projects. Some time for this work should be included.

4. Acknowledgements

Work on this paper was conducted within the *Centre for Sustainability of Linguistic Data* (*Zentrum für Nachhaltigkeit Linguistischer Daten*, NaLiDa), which is funded by the German Research Foundation (DFG) in the Scientific Library Services and Information Systems (LIS) framework, and within the infrastructure project *Heterogeneous Primary Research Data: Representation and Processing* of the Collaborative Research Centre *The Construction of Meaning: the Dynamics and Adaptivity of Linguistic Structures* (SFB 833), which is also funded by the DFG.

5. References

Barkey, R., Hinrichs, E., Hoppermann, C. Trippel, T., Zinn, C. (2011): Komponenten-basierte Metadatenschemata und Facetten-basierte Suche - Ein flexibler und universeller Ansatz. In J. Griesbaum, T. Mandl & C. Womser-Hacker (eds.), Information und Wissen: global, sozial und frei? Internationales Symposium der Informationswissenschaft (Hildesheim). Boizenburg: Verlag Werner Hülsbusch (vwh), pp. 62-73.

Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., Zinn, C. (2010): A Data Category Registry- and Component-based Metadata Framework. In Proceedings of the 7th Conference on International Language Resources and Evaluation, 19-21 May 2010, European Language Resources Association.

Deutsche Forschungsgemeinschaft (1998): Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft", Denkschrift. Weinheim: Wiley-VCH. See http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf (retrieved March 31, 2011).

International Organization of Standardization (2009): Terminology and other language and content resources - Specification of data categories and management of a Data Category Registry for language resources (ISO-12620-2009), Geneva. Go to www.isocat.org to access the registry.

Wissenschaftsrat (2011): Empfehlung zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften. Berlin: 28/01/2011. See http://www.wissenschaftsrat.de/download/archiv/10465-11.pdf (retrieved March 31, 2011).