Towards Metadata Descriptions for Multimodal Corpora of Natural Communication Data

Farina Freigang and Kirsten Bergmann
firstname.lastname@uni-bielefeld.de
Faculty of Technology, Center of Excellence "Cognitive Interaction Technology" (CITEC)
Collaborative Research Center "Alignment in Communication" (SFB 673)
Bielefeld University, P.O. Box 100 131, D-33501 Bielefeld, Germany

Abstract

Metadata play an important role for successful corpus management and reusability of corpora. For linguistic resources there already exist a large amount of metadata descriptions and metadata schemes. However, not much work has been done to develop metadata for the particular structure of *multimodal* corpora, yet. In this paper we provide a review of existing metadata profiles for multimodal data. We discuss in how far these are adequate to describe multimodal resources and point out conclusions for future efforts.

Keywords:
Metadata, Multimodal Corpora, Modality, CLARIN

1 Introduction

The production of high-quality multimodal corpora is extremely expensive and hence it is of major importance to manage these resources in a way that they are easily reusable and searchable for other researchers. In fact, the reuse of resources is an issue strongly promoted by research funding organizations, for example, by the European Union in terms of their "open data strategy". In the field of corpus linguistics and language resources it is widely agreed that the ever-expanding number and growth of corpora needs *metadata* for the purpose of corpus management, i.e., "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource" (*Understanding Metadata*, 2004, p.1). For linguistic resources there already exists a large amount of metadata descriptions and metadata schemes, but so far not much work has been done to develop metadata for the particular structure of multimodal corpora.

A prominent example among metadata schemes for linguistic resources is the Dublin Core (DC) metadata initiative[1] which originated in the library world. DC is a set of fifteen properties (e.g., title, author, date) mostly used for the description of written resources. An extension of DC is the Open Language Archive Community's metadata set (OLAC) which aimed to provide a metadata standard fitting the particular needs of language archives (Simons & Bird, 2008). Further metadata elements were added by OLAC and vocabularies of values were defined to guarantee a consistent description of language resources. At the same time as the OLAC scheme, another metadata standard has been developed by the Isle Metadata Initiative (IMDI) which aimed to define a metadata standard not only for linguistic, but also for multimedia and multimodal resources. By applying those metadata schemes to different linguistic subdomains, linguistic researchers have, however, realized that "a single metadata scheme cannot succeed in conquering all fields of linguistics" (Broeder, Windhouwer, van Uytvanck, Trippel, & Goosen, 2012, p. 1) due to major differences in needs, terminology and research traditions.

This challenge is even greater for multimodal corpora of natural communication data as we are faced with highly heterogeneous resources here: Data is collected with different research aims in mind, there is already a wide range of different kinds of primary data (video data, motion capturing data etc.). For secondary data there exist different annotation schemes of varying depth and granularity and annota-

[1] http://dublincore.org/

Freigang, F., & Bergmann, K. (2013). Towards Metadata Descriptions for Multimodal Corpora of Natural Communication Data.
In J. Edlund, D. Heylen, & P. Paggio (Eds.), Proceedings of the Workshop on Multimodal Corpora 2013: Multimodal Corpora: Beyond Audio and Video.

tions are realized on the basis of different annotation systems. So the application of a single metadata format for multimodal resources is not auspicious at all. An alternative has been proposed by Broeder, Schonefeld, Trippel, Van Uytvanck, and Witt (2011) in terms of a component-based approach. The *Component Metadata Infrastructure* (CMDI) provides a flexible framework in which users can combine several metadata components into a self-defined scheme that fits their particular needs. To this end, users can reuse existing structures and parts, but they can also define their own components and profiles. The database of those components and profiles is the Common Language Resources and Technology Infrastructure (CLARIN) Component Registry[2]. Due to its flexibility, CMDI is also able to represent other metadata schemes like DC, OLAC or IMDI.

Once components and profiles have been created and published in the CLARIN Component Registry, each element in a component needs to be uniquely defined by a persistent identifier in the CLARIN ISOcat Registry[3]. Corpora and datasets that are described by metadata profiles can be searched, for example, *via* the CLARIN Virtual Language Observatory[4] (VLO).

The goal of this paper is to analyse in how far CMDI-based metadata can adequately describe multimodal resources. In the following section we give an overview of the few existing CMDI-based attempts which aim to cover multimodal resources. These are discussed and evaluated with regard to differences, strengths and weaknesses as a basis for us to point out conclusions for the structure and realization of adequate metadata descriptions for multimodal corpora. In our discussion we focus on two issues of major importance. First, the most essential feature of a multimodal corpus is the kind of *modalities* it covers and accordingly we will pay particular attention to the term 'modality' and how metadata elements address these. Our second focus is a practical point, namely the *realization* of metadata descriptions. This is a crucial point as the idea of CMDI is that users compose self-defined metadata descriptions fitting their particular needs.

2 Profiles for Multimodal Corpora in CMDI

Existing components from which CMDI metadata can be extracted, are available *via* metadata infrastructures. The most prominent ones are CLARIN with its CLARIN Component Registry and the Multilingual Europe Technology Alliance (META) with its META-SHARE Repository (Federmann et al., 2012). One reason to work with CLARIN is that is has already a huge collection of profiles and components, including some profiles addressing multimodal resources, from which CMDIs can be generated. So far, the registry contains three profiles that aim to describe multimodal corpora: (1) the *media-corpus-profile* in combination with the *media-session-profile* from the Bavarian Archive for Speech Signals (BAS) at LMU Munich, (2) the profile *MultimodalCorpus* developed within the NaLiDa project for sustainability of linguistic data at Tübingen University, and (3) *BamdesMultimodalCorpus*, used for harvesting purposes by the Harvesting Day initiative[5]. In the following we characterize and contrast these metadata profiles and discuss them regarding their strengths and weaknesses. For an overview of the profile structures see figure 1.

Media-corpus-profile and Media-session-profile (BAS) The *media-corpus-profile* and the *media-session-profile* were first released into public space of the CLARIN Component Registry in February 2012. One year later, an updated and backward compatible version of *media-session-profile* was published, which is the version we refer to here. Used in combination, the two profiles provide a comprehensive metadata architecture, which has already been used for many corpora[6], for example, for the multimodal SmartKom Public corpus (Schiel, Steininger, & Türk, 2002). Relevant for the description of multimodal data is the component *cmdi-modality* with an element *Modality* listing modality specific descriptions: 'spoken', 'written', 'music notation', 'gestures', 'pointing-gestures', 'signs', 'eye-gaze', 'facial-expressions', 'emotional-state', 'haptic', 'song', and 'instrumental music'.

The ***media-corpus-profile*** consists of three main components: (1) *cmdi-COLLECTION*, (2) *cmdi-corpus* and (3) *cmdi-speech-corpus*. The component *cmdi-COLLECTION* provides general information about the corpus such as

[2] http://catalog.clarin.eu/ds/ComponentRegistry/

[3] https://catalog.clarin.eu/isocat/interface/index.html

[4] http://catalog.clarin.eu/vlo/

[5] http://theharvestingday.eu/

[6] For corpora referenced with the BAS metadata profiles see https://clarin.phonetik.uni-muenchen.de/BASRepository/

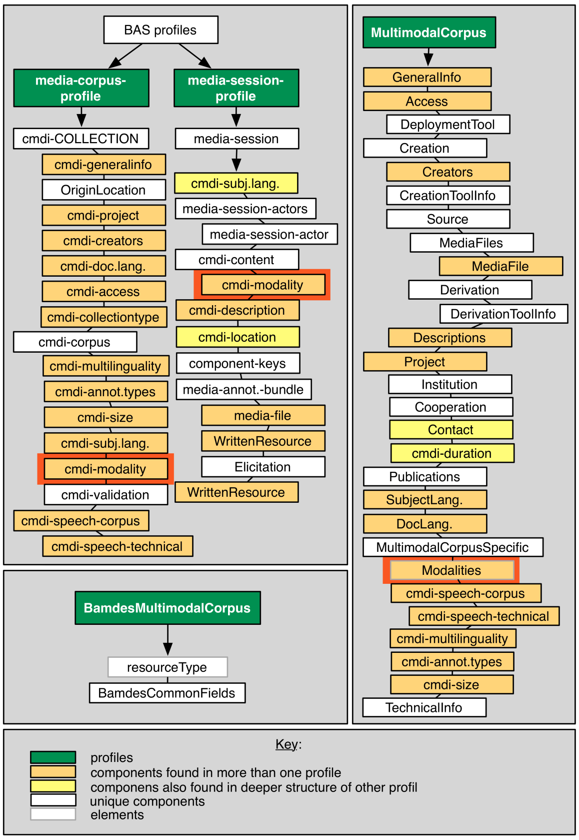

Figure 1: Three existing metadata profiles from the CLARIN Component Registry to describe multimodal corpora.

name, ID and time coverage, as well as information and contact details about the project, the creators and the access of the corpus. The component *cmdi-corpus* provides a more detailed corpus description with information about corpus size, validation and multilinguality of the corpus. Note that some information can be entered for both, the *media-corpus-profile* and the *media-session-profile*: annotation type, languages spoken during the study/sessions and corpus modalities (see above). Finally, the component *cmdi-speech-corpus* contains facts such as duration of effective speech, number of speakers and recording environment. This component can also be used for pure speech corpora if the metadata description is done without using the *media-session-profile*.

The ***media-session-profile*** is designed to describe each single session of a corpus. It consists of the component *media-session* that contains general information about recordings and environment, languages spoken during the session (‘*cmdi-subjectlanguages*’), detailed information about the participant (e.g., handedness and speech disorder) and the content of the session.

Each session can further be divided into several bundles (*media-annotation-bundle*). Bundles are useful when different tasks are performed by the same participants within an experiment. Metadata descriptions for bundles contain information about the kind of elicitation (e.g., instruction, text or medium), about the media files and the annotation files. A component *media file* provides information about the type of media, its quality, its size and facts of the recording.

MultimodalCorpus (NaLiDa) The NaLiDa profile has been released in January 2013 and is used within the Tübingen CRC 833. This profile also contains an element *Modalities*, which is almost identical to the BAS component *cmdi-modality*, but lists two more concepts, namely ‘multimodal’ and ‘transcribed’.

The overall structure is somewhat different from the BAS profiles, but the components and elements convey similar information to a large extent. The following nine components make up the main architecture of the *MultimodalCorpus*: (1) general information, (2) access, (3) creation, (4) project, (5) publications, (6) spoken language during the study (again called ”SubjectLanguages”), (7) documentation language, (8) multimodal corpus specific and (9) technical information. Further details are provided *via* further sub-compontents. For instance, the component *MultimodalCorpusSpecific* consists of the components *cmdi-speech-corpus*, *cmdi-annotation-types*, and *cmdi-size* and the element *Modalities* (see above).

BamdesMultimodalCorpus This was published in an early stage of the CLARIN Component Registry, in October 2010. BAMDES is a metadata format standing for Basic Metadata Description. The profile has been developed for ‘The Harvesting Day’ initiative, ”a metadata harvesting routine based on the OAI-PMH protocol for metadata harvesting” (Parra, Villegas, & Bel, 2010). There are no elements allowing for multimodal metadata descriptions.

	BAS	NaLiDa
Profile Structure	Profile and session components separate Media and annotation close together in the hierarchy	One profile component and no separate session components Media and annotation files in separate components
Profile Components	*media-session-actor*, *media-session*, *media-annotation-bundle*, *cmdi-validation*, *Elicitation*	*Cooperation*, *Publications*, *TechnicalInfo*, various tool information
Profile Realization	Existing components reused (e.g., from CLARIN-NL)	New components created: minor and major modifications made to many existing components

Table 1: Comparison of BAS and NaLiDa profiles.

The profile consists of several elements and one component (*BamdesCommonFields*) which consists of further elements. For example, the element *resourceType* has one value 'MultimodalCorpus', whereas *corpusType* contains many values that do not necessarily belong to one concept class and from which only one can be taken: 'monolingual', 'bilingual', 'spontaneous', 'dialogue', 'monologue' etc.

Discussion of existing Profiles Both, the BAS metadata profiles and the *MultimodalCorpus* are large profiles with a wide range of classification possibilities and, thus, they are generally suitable for multimodal corpus description. By contrast, the *BamdesMultimodalCorpus* has a shallow structure, contains only a few elements and for that reason does not allow for a detailed corpus description. The BAS and the NaLiDa profiles differ from each other with regard to three major issues, as listed in table 1.

Generally, the hierarchical structures of the profiles differ to a great extent. Whereas the BAS metadata descriptions, following the tradition of IMDI metadata modeling, contain separate profiles for the overall corpus and for single sessions, the session concept is not realized in the *MultimodalCorpus*. This is partly due to the fact that the *MultimodalCorpus* lists media files under *Creation* and annotation files under *MultimodalCorpusSpecific*, pulling two relevant components for a session wide apart in the hierarchy, while in the BAS *media-session-profile* they are gathered under one component *media-annotation-bundle*. It can be argued whether information about participants (non-existent in *MultimodalCorpus*) belong to a bundle as well or can loosely be placed within the sessions component. From the perspective of multimodal corpus research, the former option is to be preferred because a participant's multimodal behavior might differ obviously from one situation to another.

Additionally, there are differences of the metadata representation between the BAS and the NaLiDa profiles at the component level. On the one hand, the BAS profiles contain relevant attributes such as *media-session-actor* (participants) including many elements, *media-session*, *media-annotation-bundle*, *cmdi-validation* and *Elicitation*. The *MultimodalCorpus* contains other components such as *Cooperation*, *Publications*, *TechnicalInfo* (about the language scripts) and various tool information (*DeploymentTool*, *CreationToolInfo* and *DerivationToolInfo*). For a comprehensive description of multimodal resources at the level of metadata, it would be nice to have a broad set of components and elements available which would ideally cover attributes from both profiles.

With respect to the question of realization, it is conspicuous that the component usage is implemented differently within the two projects. In the BAS profiles, many existing components were reused, following the overall CMDI philosophy. For example, *media-session* is one of the profile's main components and has been created by BAS, however, many of its subcomponents (e.g., *cmdi-content*) belong to the very first components of the Component Registry developed by researchers of the project CLARIN-NL[7]. By contrast, many components were cloned and newly created in the NaLiDa project. Sometimes only minor changes were made in comparison to the original ones. For instance, the NaLiDa component *Descriptions* allows for one or more occurrences of the element *Description*, whereas in the CLARIN-NL component *cmdi-description*, the same element *Description* is restricted to exactly one. Other components, like the CLARIN-NL component *cmdi-location*, were subject to more substantial modifications: In the NaLiDa component *Location* the elements *Address* and *Region* are newly defined in the ISOcat Registry, the component *ISO-continent* has been replaced by the element *ContinentName* and component *iso-country* has been totally refurbished to the component *Country*.

[7] http://www.clarin.nl

3 Conclusions for Metadata Descriptions of Multimodal Resources

The profiles described and compared in the previous section are not fully adequate to describe multimodal resources of natural communicative data in a way that the resources are searchable. This is mainly due to the fact that a short list of modality values, as in *cmdi-modality* and *Modalities*, cannot capture the depth and diversity of multimodal data such as speech and gesture, sign language and motion capture data among others. Precise modality entries are essential, however, for a suitable multimodal data description. In this section, we try to deduce a definition of modality and possible modality values from what we see in our data and then discuss ways of realizing these metadata descriptions.

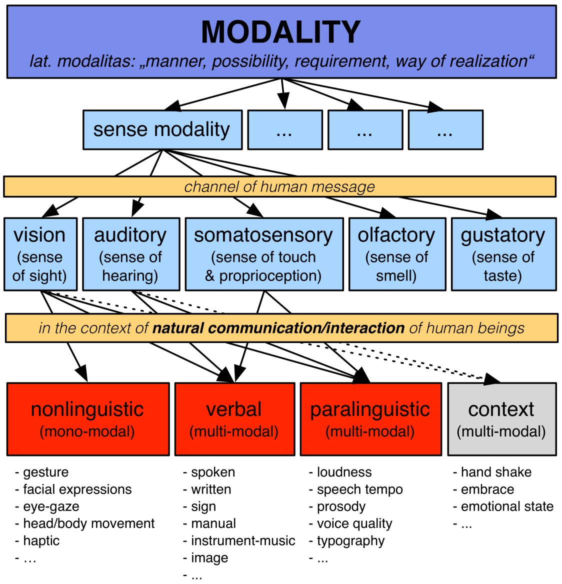

Figure 2: One way to resolve ‘modality’ for a multimodal metadata description. The values at the bottom of the figure are possible search terms used for multimodal corpora.

Modality Information First, we need to clarify the term ‘modality’. Overall, one can distinguish between the modality of a *data format* (e.g., text, audio, video, time series, etc.) and the modality of a *transmission channel* (spoken, written, sign, image, etc.). Since data formats are sufficiently described by existing metadata, we concentrate on descriptions of the transmission channel. Furthermore, it is important to differentiate the *kind of messages* covered by the multimodal corpus data in terms of verbal, nonlinguistic and paralinguistic messages which again subdivide into modality-characteristic values. The latter level is that of the modality elements in the profiles discussed above. Note, that this is just one way of presenting modalities in a hierarchy. Anyway, we go beyond plain lists of modality labels and rather reflect the deeper structure as visualized in figure 2. Along similar lines, Menke and Cimiano (2012) summarized categories for multimodal annotations, where several data units are assigned to modality categories.

Second, given the vast heterogeneity of multimodal corpora we further suggest to define *multiple components* for the different kinds of messages and sub-messages. As an example, for gestures these could be form features like ‘handedness’, ‘handshape’, or ‘palm orientation’ which might have closed or open vocabularies. It is also conceivable to have several alternative elements or components representing competing value sets. Handshapes, for example, are sometimes coded following the alphabet of the American Sign Language or following the HamNoSys notation (Prillwitz, Leven, Zienert, Hanke, & Henning, 1989).

Third, one should be able to express whether (and which) *intermodal relations* are present in the resource. It’s a particular strength of multimodal corpora that several modalities can be put in relation to each other by linking annotation elements from different tiers. For instance, gesture affiliations with words can be coded in the data. These relations are useful and important for analyses of, for example, information distribution across modalities or analyses of temporal synchronization.

Realization How could creators of multimodal corpora proceed to generate adequate metadata for their resources? One could, of course, reuse and expand the closed vocabulary lists (*Modality* or *Modalities*) to describe modalities in more detail but without a clean typological distinction. This solution is easy to realize, but not satisfying as it gives little room for precise metadata description. A second option would be to add an optional component to *cmdi-modality*/*Modalities* with modality specific elements. Usually, component structures cannot be changed once they are published, due to the persistent compatibility. If at all, only minor modifications on the structure of the component can be made in cooperation with the originator. Nevertheless, although such a modified component structure provides more description possibilities, it does not come very close to our preferred definition of modality (see above). A third option would be to define a complex component *cmdi-multimodal-specific* containing the previously mentioned structure. Such a component needs to be optional and could be listed just before or after *cmdi-modality* with the surrounding architecture staying the same. This way, one could ensure backward compatibility for the modality component. Certainly, it would mean to build a new profile but the only update would be the local component and no further thought needs to be given to the overall component architecture. Finally, another option is to create a completely new metadata profile. To this end one could rely on *cmdi-multimodal-specific* and other components already defined in the CLARIN Component Registry. However, new elements have to be defined one by one in the ISOcat Registry. A new architecture would of course perfectly fit the own corpus and hopefully many other multimodal corpora. Still, this option is difficult to realize for users not having much experience with metadata descriptions, yet.

Building upon the issues discussed here, our next step will be to set up a metadata description for the Bielefeld Speech and Gesture Alignment (SaGA) corpus (Lücking, Bergmann, Hahn, Kopp, & Rieser, 2013) developed within the Bielefeld CRC 673 'Alignment in Communication', whereby the focus will be on an extended modality component, at first. Also important is a clearly arranged component structure which offers everything needed for multimodal metadata descriptions without confusing the user. Finally, we plan to extend the components so that other multimodal corpora can also be accommodated adequately.

Acknowledgments

We thank Florian Schiel, Christoph Draxler and Bernhard Jackl for their support. Additionally, we give thanks to the anonymous reviewers for their comments and ideas. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 "Alignment in Communication", the Center of Excellence 277 "Cognitive Interaction Technology" (CITEC), and CLARIN-D, the German division of the "Common Language Resources and Technology Infrastructure".

References

Broeder, D., Schonefeld, O., Trippel, T., Van Uytvanck, D., & Witt, A. (2011). A Pragmatic Approach to XML Interoperability – the Component Metadata Infrastructure (CMDI). In *Balisage: The Markup Conference* (Vol. 7).

Broeder, D., Windhouwer, M., van Uytvanck, D., Trippel, T., & Goosen, T. (2012). CMDI: A Component Metadata Infrastructure. In *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR Workshop Programme* (pp. 1–4).

Federmann, C., Giannopoulou, I., Girardi, C., Hamon, O., Mavroeidis, D., Minutoli, S., & Schröder, M. (2012). META-SHARE v2: An Open Network of Repositories for Language Resources including Data and Tools. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC2012).*

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2013). Data-based Analysis of Speech and Gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its Applications. *Journal on Multimodal User Interfaces*, *7*(1-2), 5–18.

Menke, P., & Cimiano, P. (2012). Towards an Ontology of Categories for Multimodal Annotation. In *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR Workshop Programme* (pp. 49–54).

Parra, C., Villegas, M., & Bel, N. (2010). The basic Metadata Description (BAMDES) and theharvestingday.eu: Towards Sustainability and Visibility of LRT. In *Proceedings of Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management at LREC* (pp. 49–53).

Prillwitz, S., Leven, R., Zienert, H., Hanke, T., & Henning, J. (1989). *HamNoSys Version 2.0: Hamburg Notation System for Sign Languages: An Introduction*. Hamburg: Signum Press.

Schiel, F., Steininger, S., & Türk, U. (2002). *The SmartKom Multimodal Corpus at BAS* (Tech. Rep.). LMU Munich.

Simons, G., & Bird, S. (2008). *OLAC Metadata* (Tech. Rep.). Open Language Archive Community.

Understanding Metadata. (2004). National Information Standards Organization Press. Bethesda, MD, USA.