Requirements and approaches for generic metadata components

Hanna Hedeland
Hamburg Center for Language Corpora
Universität Hamburg, Germany
hanna.hedeland@uni-hamburg.de

Felix Rau
Department of Linguistics
University of Cologne, Germany
f.rau@uni.koeln.de

Twan Goosen
CLARIN ERIC
Utrecht, The Netherlands
twan@clarin.eu

Abstract

Modelling recommended components of basic concepts for description of language resources and tools is a major challenge for the Component MetaData Infrastructure. These components are required to be adequate, interoperable, and reusable. We describe the ongoing effort to create such an inventory of existing attempts and discuss the BLAM profiles for language archiving as one of these approaches. We further examine the challenges to model metadata for interoperability outside the LRT context. Finally, we identify the need for a standardized means of defining relations between versions of profiles and components as well as concepts across registries and integrated workflows to handle changing and varying definitions.

1 Introduction

This paper presents work in progress based on completed work as well as ongoing efforts within several contexts related to the design and implementation of a recommended *general information* component comprising itself a set of recommended components, each modelling basic concepts relevant for the description of language resources and tools. Following the compilation of CMDI Best Practices[1] by the CMDI and Metadata Curation SCCTC Taskforces (Eckart et al., 2017), this is an important task to take on. Working towards a recommended *general information* component is not merely about VLO usability or complying with best practices, but also about increased knowledge regarding a minimal set of shared metadata necessary to describe various kinds of resources and tools relevant to researchers and students within various Humanities and Social Sciences disciplines.

In the first section of this paper, we present basic requirements for the acceptance among the various stakeholders, i.e. metadata modellers, creators and users. The second section reviews previous and current work related to these requirements, while the third section elaborates on several problems encountered and possible solutions within the CMDI framework. Finally, we give some concluding remarks and hope that this paper can contribute to the general discussion on metadata best practices within CLARIN.

2 Preliminary requirements

2.1 Adequacy

For a general information component to be accepted and re-used by the various communities producing linguistic resources and tools, it needs to describe these adequately. As an example, metadata for language resources and tools will always require information on the contained or applicable ('object') languages, and a recommended metadata component therefore needs to allow for the description of all languages and varieties for which digital resources exist. In the context of language resources and tools, the description of a language for example must provide data categories that allow to identify a particular language, unequivocally and independent of the language used in the metadata. Furthermore it should facilitate discoverability across resources, by providing consistent language names so that one search term can match different resources. Ideally, it also provides more information such as alternative names

[1] Live draft version: https://www.clarin.eu/content/cmdi-best-practice-guide

and taxonomic information so that users can target different levels of linguistic classification. While univocal identification can be achieved by using norm data, language independent, unambiguous identification requires identifiers such as the codes of the ISO 639 set or Glottocode (Hammarström et al., 2018). ISO 639-3 is the most suitable ISO standard. However, the ISO 639-3 codes can be insufficient for underdescribed languages (Good and Cysouw, 2013) and Glottocodes might be best suited to identify underdescribed languages (Hammarström and Nordhoff, 2011). The Section 3.2 covers this example of language metadata more in detail.

2.2 Interoperability

CMDI allows for a wide range of existing metadata schemas to be modelled using CMD profiles and components, and coverage of widely used schemas are therefore an important aspect. Most often, CMD profile interoperability has been a question of mapping from community standards, e.g. TEI or IMDI (cf. (Hansen et al., 2014), (Hedeland and Wörner, 2012)), or from existing digital archive metadata standards such as DC and OLAC, to CMDI for integration into the CLARIN infrastructure and in particular into the VLO. Though the VLO has been tailored for language resources and tools, providing metadata to further catalogs and portals increases visibility and discoverability especially for audiences not yet familiar with the CLARIN infrastructure due to their geographic location or discipline. Apart form the work on conversion to RDF for (L)LOD (Ďurčo and Windhouwer, 2014), there have been efforts to provide metadata on language corpora via library catalogs, which we will describe further in Section 3.3. With a shared set of concepts, these kinds of mapping scenarios and workflows would become easily available to all interested metadata providers by using the recommended components, whereas in the current state, conflicting concepts and structures for basic metadata do not lend themselves well to automatic processing.

2.3 Reusability

In addition to the CMDI Best Practices guidelines, the CLARIN Metadata Curation Module(Ostojic et al., 2017) provides metadata modellers and creators with comprehensive direct feedback on specific CMDI profiles and instances. However, for the metadata modeller, no recommended profile or component exists to be used as a basis to add required specialized metadata for a certain usage scenario to. While detailed metadata, especially on a language resource, tends to be highly specific to its theoretical framework, research question and methodological aspects, more general metadata is necessary for all kinds of linguistic resources and tools to be discoverable and interpretable. General information is needed for cataloging to provide enough information on data provenance to assess resource quality. We believe that rather than recommended detailed profiles for specific scenarios, a modular approach using the fundamental component metadata concept would make it easier to include a core set of metadata suitable for any kind of linguistic resource or tool into individual specialized profiles. We start the following Section 3 by describing our inventory of such *general information* components in Section 3.1.

3 First steps towards a recommended general information component

3.1 An inventory of existing components

To gain insights into the requirements of a *general information* component, relevant components already published in the component registry were gathered within a bottom-up approach. This inventory effort was followed by a comparative analysis of existing *general information* components, their constituent components and elements, and the concepts these refer to, with the purpose of identifying overlapping information types. Thus far, 9 relevant components and profiles of this type have been selected, and 7 of these have been analyzed in detail. A preliminary classification of the encountered data types yields the following set of candidate categories: identification, description, resource class, content (language, subject, modality, genre, keywords, temporal coverage, spatial coverage), legal status, organizational context, provenance, versioning, access, technical properties, documentation. Although there is substantial variation in the naming patterns, the overall level of agreement in terms of linked concepts is high. For the majority of commonly included value types encountered there exists a concept in the CLARIN Concept Registry (CCR)[2]. Additional components and profiles will be evaluated to achieve a more complete overview to be included in the final paper.

3.2 Metadata modelling within BLAM

The Basic Language Archive Metadata (BLAM) group of CMDI profiles was designed to provide basic, but adequate, metadata for resources from language documentation and related linguistic sub-disciplines. The design goal was to maximise discoverability and to facilitate interoperability with meta-catalogs, in particular DataCite, OLAC, and VLO. During the design phase, existing components describing a language were surveyed, but no component fulfilled all requirements. Several components called *Language* are not metadata about a language as such, but about the linguistic competency of a human individual. The most widely used component describing a language, for example, is *cmdi-language* (clarin.eu:cr1:c_1271859438111). At the time of writing, *cmdi-language* is used 257 times (in 101 profiles). The structure of this component and its content are minimal as can be seen in Figure 1. The component consists of one or many language names (element *LanguageName* and one ISO 639-3 code.

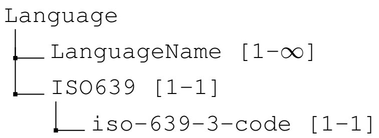

Figure 1: Component *cmdi-language*

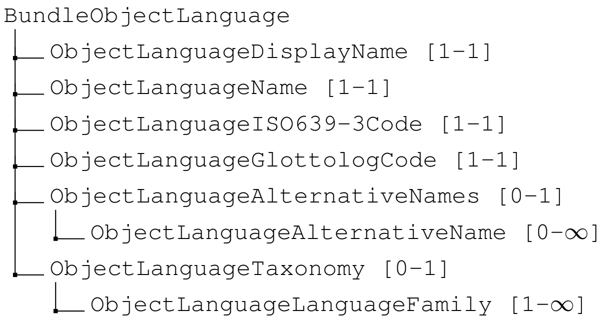
BundleObjectLanguage
ObjectLanguageDisplayName [1-1]
ObjectLanguageName [1-1]
ObjectLanguageISO639-3Code [1-1]
ObjectLanguageGlottologCode [1-1]
ObjectLanguageAlternativeNames [0-1]
ObjectLanguageAlternativeName [0-∞]
ObjectLanguageTaxonomy [0-1]
ObjectLanguageLanguageFamily [1-∞]

Figure 2: Component *BundleObjectLanguage*

Given the inconsistent nomenclature for endangered and underdescribed languages, a simple language name element is problematic for catalog consistency and discoverability. An unordered list of alternative language names with no identifiable primary name can pose problems for display and search functionalities in general and facetted searches in particular. ISO 639-3 provides an important, but insufficient identifier. For these reasons, a new (object) language component was registered for BLAM. The component *BundleObjectLanguage* has more data categories than the *cmdi-language* component. The addition of the Glottolog language code category besides the widely used ISO 639-3 codes provides a further source of language identifiers. The *ObjectLanguageName* element contains an authority controlled language name from Ethnologue or Glottolog. This ensures consistency across resources. However, these primary names may derive from derogatory exonyms and thus can be politically and socially problematic (Haspelmath, 2017), so *ObjectLanguageDisplayName* contains the name recommended by the data producer to be used for display purposes. Two additional elements contain controlled alternative names and names of the language (sub-)family to enhance discoverability of resources. While *BundleObjectLanguage* is more extensive than *cmdi-language*, six element types are not excessive, and all elements are motivated by concrete requirements to adequately describe languages in language resources.

3.3 CMDI interoperability beyond LRT

Converting from standards such as IMDI or TEI to CMDI for integration into the VLO benefits from the flexibility of CMDI and the existing language resource specific metadata as we don't move beyond the field of Language Resources and Technology (LRT). To make our resources visible beyond CLARIN and LRT, e.g. to users with a research interest only related to the content of the resource, not its linguistic features, such as Oral History related research, other contexts than the VLO might also be very useful.

The so-called "discipline specific information service for Northern Europe" (FID Nordeuropa) based at the Kiel University Library is part of a German network of such service institutions and provides in-

[2] https://concepts.clarin.eu

formation on and access to resources relevant for research related to Northern Europe. Within a pilot project, metadata for Scandinavian language corpora hosted by the CLARIN center at the Hamburg Center for Language Corpora (HZSK) was converted to library catalog standards. While the CLARIN center contributes its expertise in hosting and providing the relevant resources via its repository, the university library can create authority records and standardized metadata to be provided via library catalogs – from regional catalogs to the WorldCat. As the library standards were not created for the LRT field, mapping to concepts such as *author* or *editor* is not obvious, nor is the granularity for the description of various corpus types. However, with mappings and conversion workflows available for a general information component, re-use of the valuable resources hosted at several CLARIN centers could be increased.

4 Outlook

The inventory of existing components already showed that there is substantial overlap regarding content, but still a rather inconsistent use of similar concepts in this area. A first draft of a *general information* component based on this inventory and the further requirements discussed will be provided in the final paper. Regarding the interoperability, an unsolved issue is how to define, manage and share mappings from concepts derived from various registries or metadata standards. Currently, there is no actively managed registry for such relations, and conversion information and tools are scattered over the infrastructure. While a registry for relations might be feasible, it requires a lot of effort to manage it and keep it up to date. Another option would be to allow for mapping information to be added directly to CMD components, e.g. by allowing the use of several ConceptLinks with PIDs, URLs or other identifiers referring to various concepts registries or metadata schemas for a single component, thus shifting the responsibility to the metadata modeller and possibly allowing for generic conversion. For the first draft of a *general information* component, we will provide such mapping information within the component to be used as the basis for further discussion on a common set of metadata for the CLARIN infrastructure and beyond.

References

[Eckart et al.2017] Thomas Eckart, Twan Goosen, Susanne Haaf, Hanna Hedeland, Oddrun Ohren, Dieter van Uytvanck, and Menzo Windhouwer. 2017. Component Metadata Infrastructure Best Practices for CLARIN. In *CLARIN Annual Conference 2017 in Budapest, Hungary*.

[Good and Cysouw2013] Jeff Good and Michael Cysouw. 2013. Languoid, doculect, and glossonym: Formalizing the notion language. *Language Documentation & Conservation*, 7:331–359.

[Hammarström and Nordhoff2011] Harald Hammarström and Sebastian Nordhoff. 2011. Langdoc: Bibliographic infrastructure for linguistic typology. *Oslo Studies in Language*, 3(2):31–43.

[Hammarström et al.2018] Harald Hammarström, Sebastian Bank, Robert Forkel, and Martin Haspelmath. 2018. Glottolog 3.2. Max Planck Institute for the Science of Human History.

[Hansen et al.2014] Dorte Haltrup Hansen, Lene Offersgaard, and Sussi Olsen. 2014. Using TEI, CMDI and ISOcat in CLARIN-DK. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

[Haspelmath2017] Martin Haspelmath. 2017. Some principles for language names. *Language Documentation & Conservation*, 11:81–93.

[Hedeland and Wörner2012] Hanna Hedeland and Kai Wörner. 2012. Experiences and problems creating a cmdi profile from an existing metadata schema. In *Proceedings of LREC-Workshop "Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR"*. ELRA.

[Ostojic et al.2017] Davor Ostojic, Go Sugimoto, and Matej Ďurčo. 2017. The curation module and statistical analysis on vlo metadata quality. In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 2628 October 2016, CLARIN Common Language Resources and Technology Infrastructure*, number 136, pages 90–101. Linköping University Electronic Press, Linköpings universitet.

[Ďurčo and Windhouwer2014] Matej Ďurčo and Menzo Windhouwer. 2014. From CLARIN component metadata to linked open data. In *Proceedings of the third Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*. ELRA.