Originally published in: Digital Humanities 2011. Stanford University, Stanford, CA, USA, June 19-22, 2011. Conference Abstracts.
– Stanford: Stanford University Library, 2011. Pp. 88-90.

Trailblazing through Forests of Resources in Linguistics

Barkey, Reinhild
rbarkey@sfs.uni-tuebingen.de
Department of Linguistics, University of Tübingen

Hinrichs, Erhard
erhard.hinrichs@uni-tuebingen.de
Department of Linguistics, University of Tübingen

Hoppermann, Christina
christina.hoppermann@uni-tuebingen.de
Department of Linguistics, University of Tübingen

Trippel, Thorsten
thorsten.trippel@uni-tuebingen.de
Department of Linguistics, University of Tübingen

Zinn, Claus
claus.zinn@uni-tuebingen.de
Department of Linguistics, University of Tübingen

1. Introduction

Linguistics is facing the challenge of many other sciences as it continues to grow into increasingly complex subfields, each with its own separate or overarching branches. While linguists are certainly aware of the overall structure of the research field, they cannot follow all developments other than those of their subfields. It is thus important to help specialists but also newcomers alike to bushwhack through evolved or unknown territory of linguistic data.

A considerable amount of research data in linguistics is described with metadata. While studies described and published in archived journals and conference proceedings receive a quite homogeneous set of metadata tags — e.g., *author*, *title*, *publisher* —, this does not hold for the empirical data and analyses that underlie such studies. Moreover, lexicons, grammars, experimental data, and other types of resources come in different forms; and to make things worse, their description in terms of metadata is also not uniform, if existing at all.

These problems are well-known and there are now a number of international initiatives — e.g., CLARIN, FlareNet, MetaNet, DARIAH — to build infrastructures for managing linguistic resources. The NaLiDa project, funded by the German Research Foundation, aims at facilitating the management and access to linguistic resources originating from German research institutions. In cooperation with the German SFB 833 research center, we are developing a combination of *faceted* and *full-text search* to give *integrated* access through heterogeneous metadata sets. Our approach is supported by a central registry for metadata field descriptors, and a component repository for structured groups of data categories as larger building blocks.

2. State of Affairs

An increasing number of research institutions in linguistics is systematically archiving research data and making such data publicly available. Users can access such archives via institution-specific websites or purpose-built software where resources can be searched and, in part, downloaded. Some institutions provide access to their archives via OAI-PMH so that the archives' public content can be harvested and fed into the metadata services of data centers in the community.

The metadata provided by the various institutions differ not only in quantity and quality but also in the format of description, as Fig. 1 illustrates. Typically, a research organization designs a metadata schema that it deems to serve best its institutional setting and the number and types of resources it hosts. Different organizations will likely yield various schemas with their own structure and terminology for the schemas' nodes. Such semantic heterogeneity may be complemented by syntactic heterogeneity as formats may vary (e.g., ASCII, relational database format, XML). An organization's resources can thus be seen as a forest of trees of the same kind, whereas a tree — i.e., a description of one resource — may have deformations at the leafs as the result of not adhering to the schema. Moreover, trees will look rather naked when resources are described sparsely.

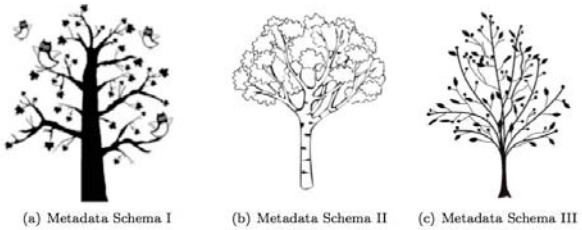

Figure 1: Metadata heterogeneity

3. Conceptual Setting

While we expect research organizations to continue managing their research data in their respective ways, we ask them (i) to make their data public

in a well-defined XML format to obtain syntactic uniformity, and (ii) to reformulate their schemas to adhere to CMDI (Component MetaData Infrastructure), see (Broeder et al. 2010), a component-based metadata model that makes use of predefined metadata components and the ISOcat data category registry (International Organization of Standardization 2009). The organizations' archive managers have the opportunity to redefine existing parts of their schema, but they can also choose to keep existing structures and terminologies. In this case the respective metadata descriptors need to be associated with their corresponding semantic points of references in ISOcat, being addressable via unique persistent identifiers. Moreover, archive managers can add new data categories to ISOcat's private space for immediate availability, and initiate a standardization process to pave the way for their wider use.

4. Technological Setting

4.1. Metadata Storage

The storage of resource descriptions has to cope with a multitude of schemas the descriptions adhere to. The use of a relational database would require a mapping of all schemas to a single one, which is all but trivial. Instead, with CouchDB [`http://couchdb.apache.org`], a no-SQL database is being used that stores arbitrarily structured documents rather than records of some fixed form. The translation between the XML-based CMDI-format into CouchDB's native JSON format is structure-and information-preserving.

4.2. Faceted Search

Facets serve to blaze the trail. Faceted search enables a user to find specific trees, i.e., resource descriptions, in the various forests by specifying (some of) their common properties. A facet partitions the search space where descriptions, i.e., CouchDB documents, in the same cluster share the same facet value. The selection of multiple facets corresponds to an intersection of clusters identifying resources that have all selected facet values. Faceted search also supplies a user with information about the number of documents in each cluster or intersections thereof — the "mileage" of following a trail. In the presently available data the following *unconditional* facets are adequate for the various schemas describing linguistic resources: "organization", "modality", "language", "country", "resourceType", and "origin". Facet values may stem from open or controlled vocabularies; controlled vocabulary facets have a stronger tendency to partition the search space into larger units, whereas open vocabularies induce larger search space fragmentations.

Once faceted search has focused on a subset of resources, *conditional* facets allow the introduction of additional context-specific navigational user aids. When users select the facet "resourceType" with value "tool", for instance, they restrict the search space to just encompass metadata that describes language-processing tools. Here, the conditional facet "toolType" is introduced that partitions the remaining search space according to the type of tool, e.g., language parser, spell checker. Moreover, conditional facets help lowering the complexity of computing search space clusters and their intersections.

4.3. Mapping Facets to Nodes of the Various Schemas

Metadata schemas vary in structure and terminology. Different names for nodes or leafs may be used to elicit the same meaning, and identical names may be used for semantically different concepts. This makes the mapping of facets to nodes of the various tree types (cf. Fig. 2) rather difficult, and usually requires some intricate knowledge of the metadata forests to be processed. When schemas make use of the aforementioned data category registry ISOcat, such ambiguities can be resolved automatically as names are linked to registry entries.

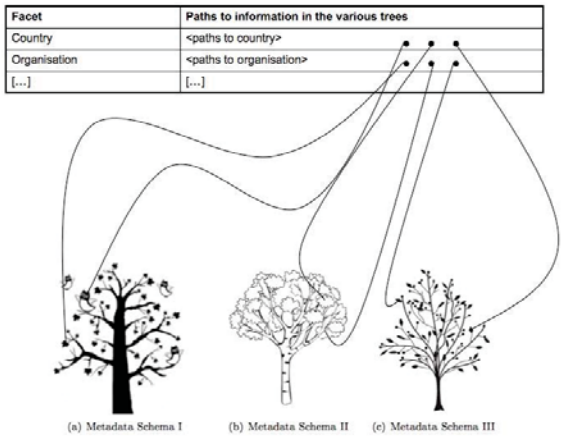

Figure 2: Mapping facets to the individual parts of the resource trees

Each facet corresponds to an elementary CouchDB view into the database of resource documents. These views serve as a starting point for the generation of complex views that correspond to the various possible navigational paths using facet selection, thus implementing the faceted browser back-end.

Elementary views and complex views are generated automatically from a facet specification file, an enriched textual encoding of the table in Fig. 2.

4.4. Full-text Search Support

Data sets may have resources that are hard to find using faceted search alone. This is true for resources with sparse metadata, or with descriptors that can rarely be mapped to facets. We are therefore using CouchDB's port to Lucene to perform full-text search across all resources or search spaces restricted by prior faceted search.

4.5. Front-end

Fig. 3 depicts a screenshot of the NaLiDa faceted browser; it shows a search state where users selected three facets ("country", "modality", "resourceType") and where the system displays an overview of the remaining search space in terms of the facets, the number of resources available for each of their values, and access to all documents selected so far.

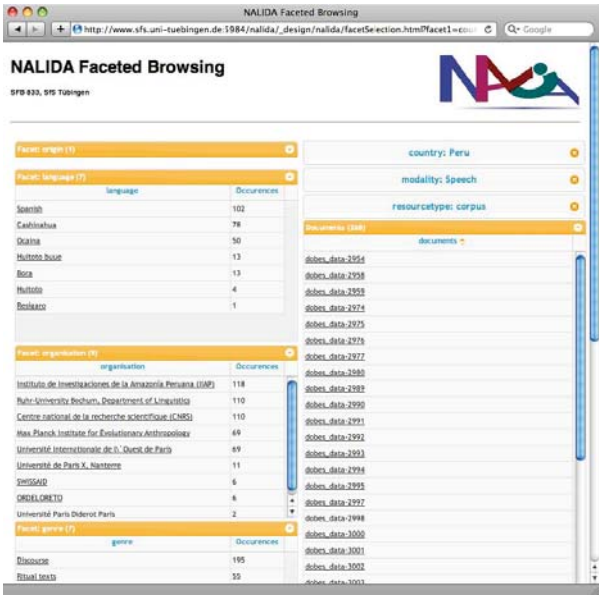

Figure 3: The NaLiDa Faceted Browser

5. Related Work and Conclusion

Faceted search is gaining popularity as users can explore large data sets without an intricate understanding of metadata fields or schemas; they obtain an immediate overview of the search space and guidance how to conquer it. A faceted search access to language resources has been implemened by the last author [`http://www.clarin.eu/vlo`] using Flamenco (Hearst 2006). Our new approach has four main advantages: CouchDB also stores the metadata documents (with varying schemas) and thus also serves as permanent storage; the use of conditional facets contributes to usability as only relevant facets are shown, guiding users' navigation; index generation accommodates for incremental updates on the metadata sets, supporting regular harvesting without recomputing all indices and views anew; and the faceted browser's back-end is generated automatically from a facet specification and can be configured easily for other datasets.

References

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'. *Proceedings of the 7th conference on International Language Resources and Evaluation, 19-21 May 2010.* European Language Resources Association.

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.* Geneva. `http://www.isocat.org`.

Hearst, M (2006). 'Design Recommendations for Hierarchical Faceted Search Interfaces'. *ACM SIGIR Workshop on Faceted Search*.