Supporting Serendipitous and Focused Search

Junte Zhang
Meertens Institute, Royal Netherlands Academy of Arts and Sciences
Amsterdam, the Netherlands

ABSTRACT

People with complex information needs are for example Humanities researchers, who need advanced search engines to investigate their research questions. Much can be gained by combining research datasets, reusing tools and serendipitously discovering new insights for further research. Humanities researchers have different (large-scale) research datasets and tools, which are described differently with metadata.

We present a highly interactive advanced search engine for Humanities researchers that semantically converges differently structured metadata records from different collections and institutions. It has features that support serendipitous and focused search in context based on the structure of the metadata used. This single system serves Humanities researchers by allowing them to search interactively across yet unexplored (research) data, discover patterns, locate relevant data for new insights, and find existing tools that could provide novel use cases.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.3.7 [**Digital Libraries**]: Systems issues, user issues; H.5.2 [**Information interfaces and presentation**]: Graphical user interfaces (GUI)

General Terms

Design, Human Factors

Keywords

information retrieval, metadata, user interfaces, ehumanities

1. INTRODUCTION

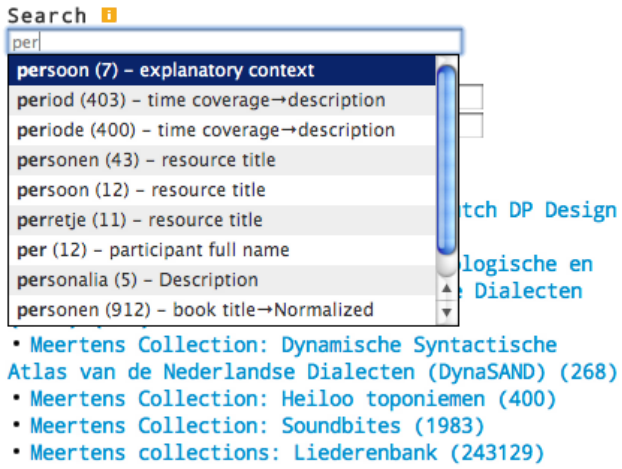
The Common Language Resources and Technology Infrastructure (CLARIN) initiative seeks to establish an integrated and interoperable research infrastructure of language resources and its technology.[1] Descriptive metadata is used to characterize large number of (legacy) research data resources (collections) and tools (e.g. Web services) to facilitate their management and discovery. The Search & Develop (S&D) project within CLARIN in the Netherlands uses the Component MetaData Infrastructure (CMDI; [4]) with ISOcat [6, 12] to open up the sharing of resources and Web services for people and machines first within the collections of a single institution, then across institutions in the Netherlands and eventually across Europe as whole. This infrastructure enables new research methods in language research and stimulates the Digital Humanities, where new insights can be gained by combining and reusing resources from different institutions and domains, and existing tools can be more effectively found and reused based on new insights.

How to use the CMDI framework with ISOcat to search for data and services, which can be understood by both people from varying disciplines and machines? The challenge is that the data is heterogenous both in content and structure, and can be massive in amount. In [11], we show how to deal with such heterogeneously structured data in the CMDI MI Search Engine. Users of the CMDI framework are mostly Humanities researchers. What type of system is needed driven by CMDI that matches with the search behavior of these users? This paper presents a proposition that has been implemented on a live system.

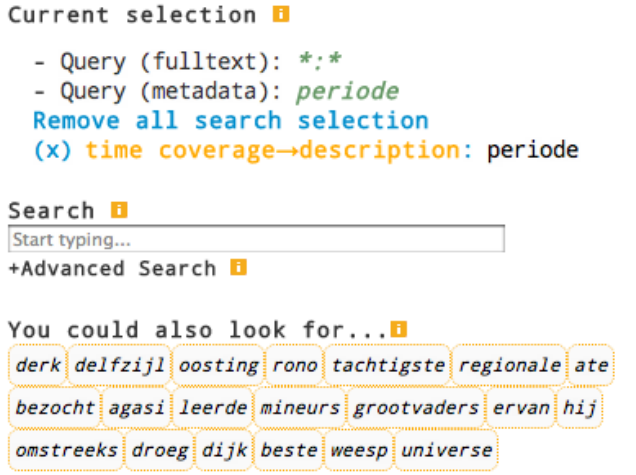
2. USING CMDI FOR FOCUSED AND SEMANTIC ACCESS

CMDI has grown out of the need to facilitate access, reuse, and interoperability using metadata [4]. A CMDI file in XML consists of a `<Header>`, `<Resources>`, and `<Components>`. The former two are fixed in structure, while the content and structure within `<Components>` is flexible and can encapsulate any data in any structured form. An XML schema can be used to make CMDI files coherent in structure for a (sub)collection and it contains references to ISOcat data categories (DC) stored in the Registry (DCR; [7, 6]). The DCR was established by the *ISO Technical Committee 37, Terminology and other language and content resources* based on the ISO 12620:2009 standard. Because multiple elements may refer to the same DC, semantic interoperability can be achieved across different datasets. A specification using the DCR and projected for example in an XML schema is called a metadata *profile* and can be (re)used for describ-

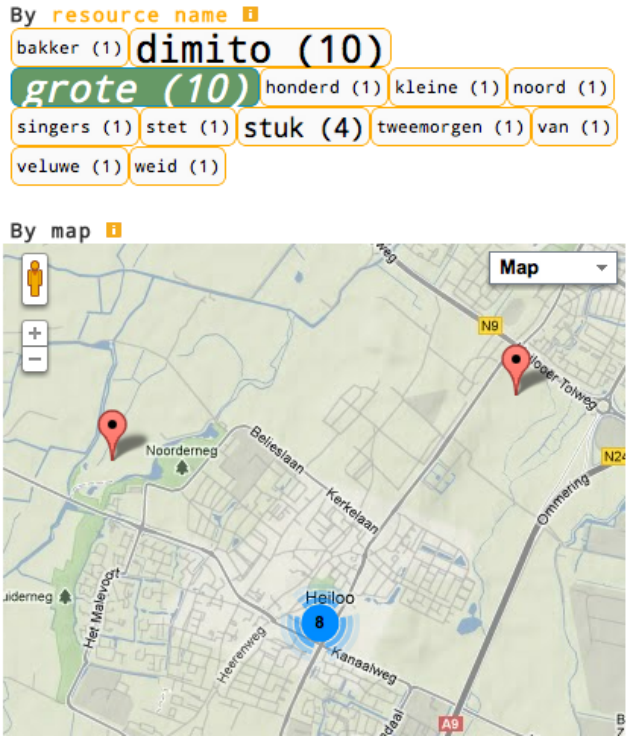
[1]See http://www.clarin.eu/external/index.php?page=about-clarin



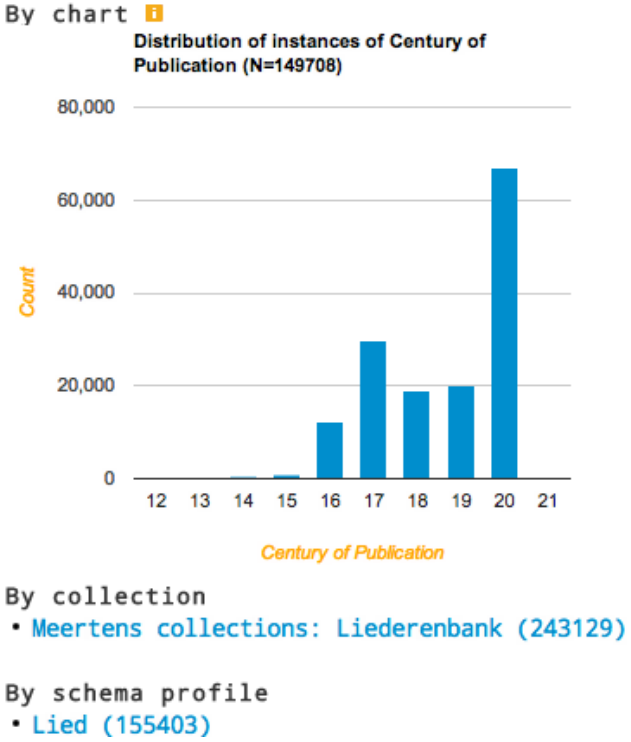
(a) Query autocompletion based on the count that a query occurs in a tag within the result set. By default the query box is content-centric, but searching directly in a tag is possible with Advanced Search (can be collapsed with a click). Users can express queries using the metadata or only the fulltext of the document by discarding autocompletion.



(b) The selection widget that allows users to keep overview of the search trail and change it, while updating the result list. Here, the query stored is "periode" (*period*) within the tag *time coverage→description*. Interesting terms are suggested by presenting the top TF∗IDF terms, which people can use to start a parallel search episode.

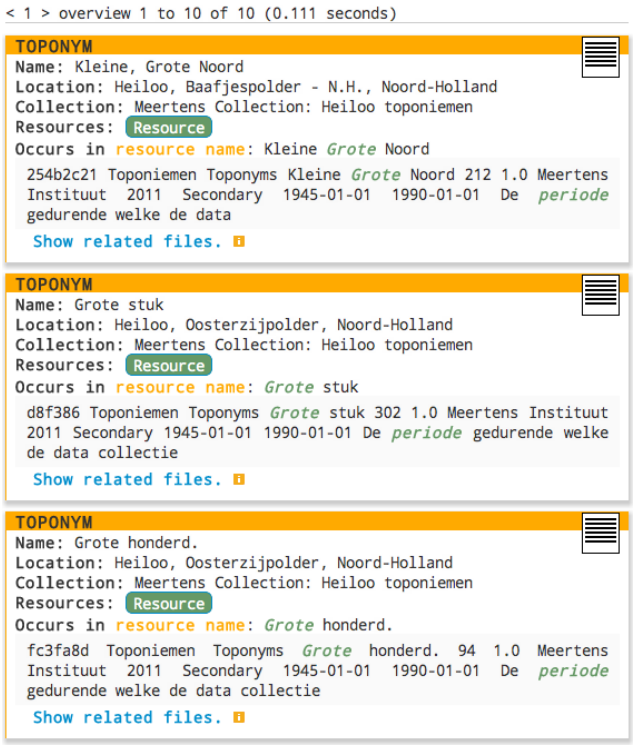


(c) To further support query expansion and serendipitous information seeking, a dynamic tag cloud is generated based on the last retrieved result list and used metadata label with keyword highlighting. Moreover, retrieved geo-referenced documents are projected on a map and clustered by markers.

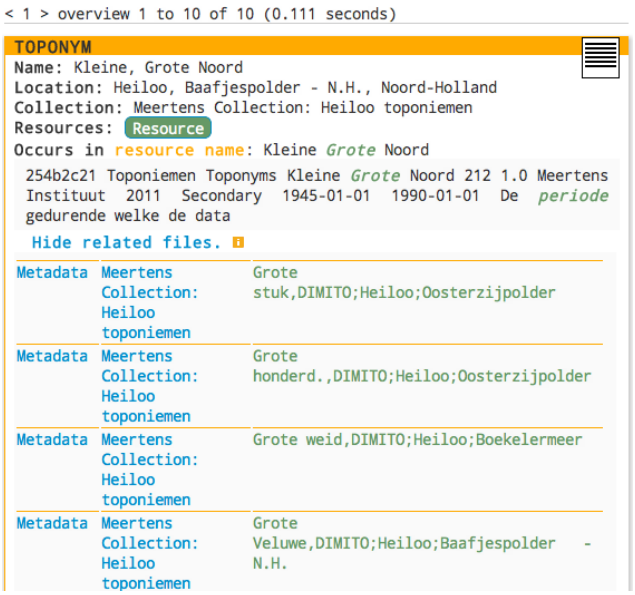


(d) The distribution of retrieved time-referenced documents (given the tags *Century of Publication* and *Year of Publication*) are visualized in bar or line charts. Users can click in the charts to narrow down the result set. The distribution of results in tags *collection* and *schema profile* always appear.

Figure 1: The CMDI MI Search Engine (1).



(a) Retrieved list of results with the display of the list of results with 'fixed' contextual information, snippets and keywords in context within the last searched metadata label and the presentation of all used keywords in context given the fulltext. There is links to the fulltext of the metadata record and the actual resource in the digital archive.



(b) For each retrieved result in the list, there is a recommendation (when available) of related results based on the content similarity of the last used metadata label. A recommendation consist of a link to the record, the collection it belongs to, and a snippet (can be collapsed with a click).

Figure 2: The CMDI MI Search Engine (2).

ing datasets and for eventual access. Moreover, RELcat [10] goes a step further by allowing for the storage of arbitrary relationships between data categories to assist crosswalks and to specify ontological relationships for further semantic search, which in the future can be used in the CMDI MI Search Engine using field collapsing.

We have indexed 246,728 CMDI files from 18 different profiles consisting of 143 different types of elements in a single stream, which shows our indexing method for CMDI files is robust enough to deal with complex data [11]. By indexing metadata in CMDI on the XML element level, the search engine can provide focused access [8]. We use straight-forward information retrieval techniques only. The 'Liederenbank' (*Dutch Song Database*) alone has 9 different profiles (XML schemas), which is equivalent to a sub-collection, ranging from very differently structured descriptions about songs to singers. How to provide interactive access to such heterogeneously structured data for Humanities researchers?

3. SERENDIPITY IN CONTEXT

When a user with no a priori intentions interacts with a node of information and acquires useful information, then serendipitous information retrieval occurs [9]. The success of serendipitous discovery is not just the find itself, but being able or willing to do something with it, so that users get more insight and can enhance the domain expertise [1]. Humanities researchers are the type of users who can be greatly supported in their research tasks with serendipitous IR, because their information-seeking behavior can be described as an idiosyncratic process of constant reading, "digging," searching, and following leads [2]. This confirms with the Berrypicking model of [3], such as that queries are not static, but rather evolve, and users "gather information in bits and pieces instead of in one grand best retrieved set."

Since the CMDI MI Search Engine should serve Humanities researchers, we design it to support serendipitous search and be highly interactive. The system has been designed to maximize the user's ability to explore. This is our focus. The user interface of the system is depicted in Fig 1. It uses the JavaScript library AJAX Solr[2], which has been heavily modified and extended by us with JQuery. It allows for faceted search [5] as we treat the indexed elements of the CMDI files as one large category hierarchy.

A user can improving the search episode (session) by effectively reducing the information space step by step. These steps are stored as part of the search trail, so the overview is kept. There are different search strategies possible. Users can search by fulltext by entering a query. This makes sure users can always search in everything. The query get highlighted in context given the fulltext, but the dynamic tag cloud widget that supports query expansion is not activated, see Fig.1(a). Users can also do a focused search request by using structure, i.e. within the content of a specified tag, and get the content of these tags returned. This can be content-centered, as users enter a keyword and the autocompletion widget returns a list consisting of keyword plus field name and hit count. It can also be structure-centered (using the Advanced Search option) by looking up a tag and then entering a keyword also with the autocompletion feature. When the last two options are used, then the keyword highlighting also occurs within the context of the retrieved

[2]See https://github.com/evolvingweb/ajax-solr

snippets of the searched tag, see Fig.2(a).

A challenge is how we can support serendipitous search given the diversely structured metadata in CMDI. Hence, we introduce and propose the concept of serendipitous search in context. We can use the heterogeneous structure of different collections to provide context to the user in a single search engine. We propose the following contextual system features that aim to support serendipitous and focused search.

- Help users by automatically completing the query that the user is entering while simultaneously and directly giving the hit count for the suggested queries in conjunction with a tag, see Fig.1(a).
- Provide inline suggestions (*Did you mean...*) based on a spell checker whenever applicable.
- Suggest a new parallel search episode (*You could also look for...*) by presenting interesting terms based on the content of the first few retrieved results after each used query, see Fig.1(b). This increments and becomes more focused as a search episode gets more queries.
- Offer different overviews of the retrieved results and allow for query expansion by directly presenting a dynamic tag cloud of the aggregated content within the metadata label used and highlighting the query entered in this context, see Fig.1(c).
- Preserve the overview of a search episode by storing the search selection (see Fig.1(b)), and the overview on collection level by the result type, e.g. the metadata profile 'lied' (*song*) in the Dutch Song Database, and the collection a document belongs to (see Fig.1(d)).
- Aggregate and visualize collection-specific search features in extra widgets, such as projecting and clustering the list of retrieved geo-referenced resources on a map (see Fig. 1(c)), and displaying the date ranges of the documents in charts that can be clicked to narrow down a result set (see Fig. 1(d)).
- Entice users to explore further by recommending related resources using the content similarity by presenting a link to the metadata record and a snippet of a recommendation, see Fig.2(b).

So the context consists of different modalities and features existing in the structure of the metadata of a collection, and used in the retrieval and visualization of information. This can be displayed on a aggregated level based on the set of retrieved results. And it can be displayed with different displays of the result types given the metadata profile. Eventually, the user finds the links to the resources in the digital archive using the metadata, and can use the found resources for further research or development. However, there is no real definite end of the search episode as people still can continue searching using the above proposed system features.

4. CONCLUSIONS

We have presented a working proposition for serendipitous and focused search by describing the CMDI MI search engine. The novelty is that it provides semantic access to diversely structured language and digital heritage resources with different metadata schemas for users such as researchers with very specific and complex information (research) needs. The search engine provides faceted search and has serendipitous features that maximize the user's ability to explore any metadata in CMDI in context, such as query autocompletion, tag clouds, and recommendation of related resources, while keeping track of the search trail. It is a tool that provides interactive and focused access to heterogeneous metadata, gives new perspectives on legacy (research) data and tools, and provides new insights for research and development. It has been released as live, and can be used at www.meertens.knaw.nl/cmdi/search.

5. ACKNOWLEDGMENTS

This work is part of the Search & Develop project at the Meertens Institute, and funded by CLARIN-NL.

6. REFERENCES

[1] P. André, m. schraefel, J. Teevan, and S. T. Dumais. Discovery is never by chance: designing for (un)serendipity. In *Proceedings of the seventh ACM conference on Creativity and cognition*, C&C '09, pages 305–314, New York, NY, USA, 2009. ACM.

[2] A. Barrett. The information-seeking habits of graduate student researchers in the humanities. *The Journal of Academic Librarianship*, 31(4):324 – 331, 2005.

[3] M. J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424, 1989.

[4] D. Broeder, M. Kemps-Snijders, D. V. Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. A data category registry- and component-based metadata framework. In *LREC*, 2010.

[5] M. A. Hearst and C. Karadi. Cat-a-cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In *SIGIR*, pages 246–255, New York, NY, USA, 1997. ACM.

[6] M. Kemps-Snijders, M. Windhouwer, P. Wittenburg, and S. E. Wright. ISOcat: remodelling metadata for language resources. *IJMSO*, 4(4):261–276, 2009.

[7] M. Kemps-Snijders, C. Zinn, J. Ringersma, and M. Windhouwer. Ensuring semantic interoperability on lexical resources. In *LREC*, 2008.

[8] M. Lalmas. *XML Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.

[9] E. G. Toms. Serendipitous information retrieval. In *DELOS Workshop: Information Seeking, Searching and Querying in Digital Libraries*, 2000.

[10] M. Windhouwer. RELcat: a relation registry for isocat data categories. In *LREC*, 2012.

[11] J. Zhang, M. Kemps-Snijders, and H. Bennis. The CMDI MI Search Engine: Access to language resources and tools using heterogeneous metadata schemas. In *TPDL*, volume 7489 of *Lecture Notes in Computer Science*. Springer, 2012.

[12] C. Zinn, C. Hoppermann, and T. Trippel. The isocat registry reloaded. In *The Semantic Web: Research and Applications*, volume 7295 of *Lecture Notes in Computer Science*, pages 285–299. Springer Berlin / Heidelberg, 2012.