MTAS – Extending Solr into a Scalable Search Solution and Analysis Tool on Multi-Tier Annotated Text

by Matthijs Brouwer (Meertens Institute, KNAW)

To deliver searchability on huge amounts of textual resources and associated metadata, the Lucene based Apache Solr index provides a well proven and scalable solution. Within the field of Humanities, textual data is often enriched with structures like part-of-speech, named entities, sentences or paragraphs. To include and use these annotations in search conditions and results, we developed a plugin that extends existing Solr functionalities with search and analysis options on these layers.

The Lucene approach to process textual data is based upon tokenisation of the provided information: for each document and field, the occurring tokens or words are stored within the index together with positional information and an optional payload. This offers quick retrieval of matching documents when searching for specific words, and specific sequences can be found by efficiently exploiting positional information on only those documents that contain all words involved. Mtas extends this technique by encoding a prefix and postfix within the value of each token. Since Lucene allows the use of multiple tokens on the same position, this provides the capability to store multiple layers of single positioned tokens, where layers can be distinguished by the applied prefix. Furthermore, to process non-single positioned and hierarchical related elements, additional information is stored within the payload, allowing these items to be stored as single positioned tokens on their first occurring position in the Lucene index structure. Finally, to enable fast retrieval of information based on position or hierarchical relation within the document, forward indices are created.

Although the default Lucene query mechanism can still be applied, specific methods are needed and made available within Mtas to efficiently use the additional encoded information and indices. Based on these methods, within Mtas a parser is provided for the Corpus Query Language (CQL), making it possible for users to define advanced conditions on the annotated text directly in the Solr search request. Since existing Solr functionality is maintained, this can be combined with regular defined conditions on metadata fields within the same document. For selected documents, Mtas can efficiently generate frequency lists for occurring layers, provide statistics over matches for possibly multiple user defined CQL queries, categorize these by one or multiple metadata fields, produce keyword-in-context representations, and group results over one or multiple layers. Mtas fully supports the distributed search capabilities from Solr, providing not only scalability but also making the process of updating and extending the data with new collections easier.

Parsers are available for multiple often XML based annotated document types, e.g. FoLiA, ISO-TEI and CHAT. The mapping by the parser of these usually

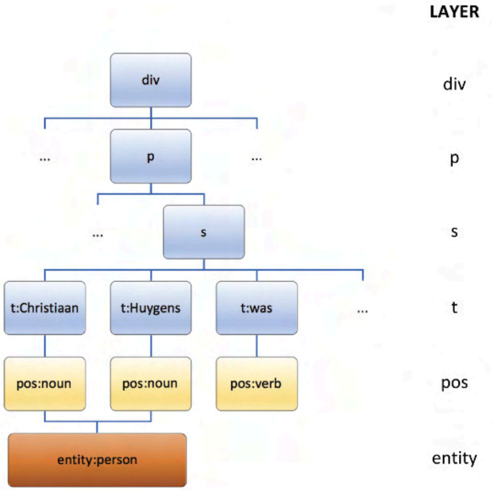


Figure 1: Typical structure of textual data with relations and annotations on multiple tiers as can be processed by Mtas.

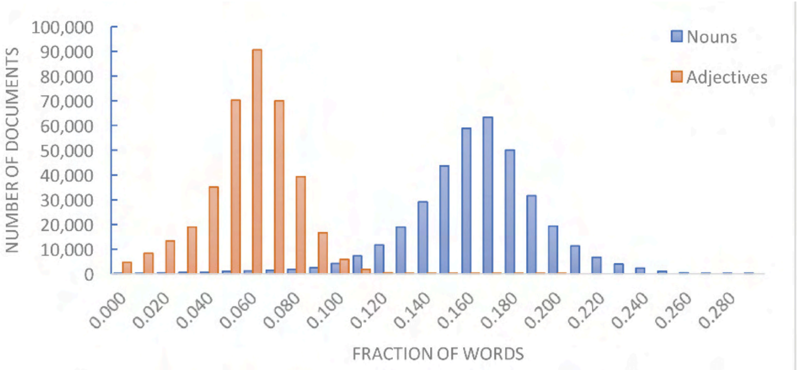
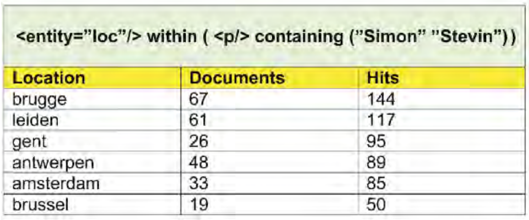


Figure 2: Distribution of the fraction of nouns and adjectives over all documents in the Nederlab collection with at least 500 part-of-speech annotated words.

Figure 3: Location entities in paragraphs containing 'Simon Stevin', constructed with Mtas by grouping all matches on a CQL expression.



<entity="loc"/> within (<p/> containing ("Simon" "Stevin"))

Location	Documents	Hits
brugge	67	144
leiden	61	117
gent	26	95
antwerpen	48	89
amsterdam	33	85
brussel	19	50

somewhat loosely defined formats onto the index can be configured in detail to include in a coherent way (only) those layers that are of interest. Multiple documents and multiple mapping configurations can be combined within the same index and field. Source code, documentation, example configurations and a Docker demonstration version are available from GitHub [L1].

The Nederlab project [L2] is one of the primary use cases for the system. It aims to bring together all digitized texts relevant to the national heritage of the Netherlands, the history of Dutch language and culture (c. 800 – present) in one user-friendly and tool-enriched open access web interface, allowing scholars to simultaneously search and analyse data from texts spanning the full recorded history of the Netherlands, its language and culture. It currently provides access to over 15 million items or documents, containing almost 10 billion positions or words and over 35 billion annotations. This data is distributed over 23 different cores with a total index size from over one terabyte, and hosted on a single Xeon E5 server with 128 GB memory. Query time, of course depending on the type of request, is usually in the order of seconds and limited by server configuration to three minutes.

Some experiments on the use of topic modelling techniques with results from Mtas have been performed. This was heavily based on the expensive generation of frequency lists, and subsequent outcomes are not used to refine or extend analysis and search within Mtas. We would like to improve the efficiency of applying these techniques and offer methods to use results again for further research within conditions on the Solr index. Another ambition involves adjustments of the Mtas index structure to incorporate frequency lists on multiple layers (e.g., to create a list of all used adjectives), which could also improve performance for certain queries. Furthermore, we are interested in implementing an additional query language to explore the stored hierarchical structure, since this is currently not covered very well by CQL. Finally, to better support and integrate sequence-based techniques, it will probably be necessary to extend the index structure with n-grams, whilst also improving certain types of already supported queries.

Links:
[L1] https://kwz.me/hmd
[L2] https://kwz.me/hmc

Reference:

[1] M. Brouwer, H. Brugman, M. Kemps-Snijders, MTAS: A Solr/Lucene based Multi-Tier Annotation Search solution, Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence

Please contact:

Matthijs Brouwer
Meertens Institute, KNAW, Amsterdam, The Netherlands
matthijs.brouwer@meertens.knaw.nl

Phonetic Search in Audio and Video Recordings

by Ioannis Dologlou and Stelios Bakamidis (RC ATHENA)

A new system uses advanced speech recognition technology to easily and efficiently retrieve information from audio/video recordings just by using keywords.

The massive amount of information produced by today's media (radio, television, etc.) and telecommunications (fixed, mobile telephony, satellite communications, etc) necessitates the use of automatic management strategies. Useful information can be retrieved from audio/video files by using keywords, in the same way as for text files, with a system that automatically searches for appropriate information in audio/video files using a state-of-the-art voice recognition engine. This enables valuable information in broadcast news or telephone conversations to be retrieved easily, quickly and accurately.

This system was developed by Voice-In SA, a spin-off company of the Greek Research Centre RC ATHENA [L1]. The research started in 2008 and the first system was delivered two years later. Research is ongoing to improve the performance and speed of the algorithms involved.

The proposed system implements the most advanced speech recognition technology (large vocabulary, continuous speech, speaker independent). It converts the statistical models of the speech recognition system and adapts them to increase both flexibility and efficiency over the handling of information which is provided by the keywords. In addition the new approach comprises a scoring algorithm for automatic detection of words or phrases that are closest to the user's query.

The system consists of two subsystems. The first subsystem performs a pre-processing on each new archiving material (recordings or video files), so that a file with specific information is created. The second subsystem is the actual core of the system that implements the new algorithms for search and retrieval, simultaneously exploiting the previously stored information.

The input to the system is audio or video files along with some keywords that the user wants to locate in these files. Following a very fast processing of the input data, the system provides information on whether the keywords are present in those files or not. If the outcome of the search is positive, the specific audio or video spots that have been found are mined and supplied to the user accompanied by the exact