Beyond Metadata – Enriching Life Science Publications in *LIVIVO* with Semantic Entities from the Linked Data Cloud

Bernd Müller*
ZB MED Leibniz Information Centre for Life Sciences
Gleueler Str. 60
50937 Cologne, Germany
bernd.mueller@zbmed.de

Alexandra Hagelstein
ZB MED Leibniz Information Centre for Life Sciences
Gleueler Str. 60
50937 Cologne, Germany
hagelstein@zbmed.de

ABSTRACT

Queries in literature search engines are usually conducted on metadata derived from scientific publications. The search engine LIVIVO holds a corpus of 63 Million life science publications. About 25 Million publications in LIVIVO are taken from PubMed that have annotations with Medical Subject Headings (MeSH). The other publications have heterogeneous keyword annotations. Hence, a workflow is developed using the Unstructured Information Management Architecture (UIMA) to enrich publications from LIVIVO with semantic annotations. The UIMA analysis engine ConceptMapper employs entity recognition based on dictionaries developed using MeSH, the pharmaceutical database DrugBank, and the multilingual agricultural vocabulary AGROVOC . Additionally, ontological relationships amongst the semantic entities are preserved by using the graph database Neo4j. The ontological information is derived from the MeSH tree, the Anatomical Therapeutic Chemical classification system (ATC) for pharmaceuticals and the AGROVOC tree. The ontological structure of semantic entities enables functionalities like query expansion, the aggregation of search results, and concept-based ranking algorithms.

Demo: http://labs.livivo.de
JSON-LD: https://datahub.io/dataset/livtdm

Keywords

linked data, graph database, document database, named entity recognition, semantic search

1. INTRODUCTION

The rapid growth of scientific literature necessitates an automated analysis of the unstructured information content. The total number of references in the literature database Medline exceeds 25 Million citations with a 4% growth per year [6]. LIVIVO[1], Europe's largest literature search engine for life sciences maintained by ZB MED[2] Leibniz Information Centre for Life Sciences, holds currently about 63 Million citations including the corpus from Medline as well as scientific publications in various languages within distinct areas of research in the fields of medicine, health, nutritional, environmental, and agricultural sciences. The 63 Million citations include 34 Million English publications, 8.8 Million German publications, 1.5 Million French publications, and 1 Million Spanish publications.

The metadata of publications helps to search and categorize documents based on their keyword annotations like MeSH on Medline citations. Keywords are often assigned by professional human indexers, alternatively the authors specify keywords for the publication manually. These tasks make the overall process costly and time-consuming. Therefore, an automated procedure to annotate entities in unstructured text content helps to search and find relevant information for researchers.

A UIMA [2] based text and data mining workflow is presented that automatically annotates named entities from the linked data cloud in LIVIVO's life science publications. The exploration of the possibilities of automatically classifying publications is conducted based on the named entities found in the text. State-of-the-art techniques of lexical acquisition are successfully applied to acquire terminological information from publications. This empowers large scale data analysis and visualization for three novel information retrieval functionalities in LIVIVO:

- Query expansion by detecting semantic entities with their synonyms in search terms and documents
- Aggregation of search results on the basis of co-occurring semantic entities detected in the search results.
- Concept-based ranking algorithms.

The workflow enables statistical data analysis and visualization by combining standoff annotations of the semantic entities in publications in conjunction with the relational properties of the entities.

2. METHODOLOGY

Semantic entities from the linked data cloud are taken for the generation of dictionaries that are used in a UIMA-based workflow for named entity recognition. The metadata of the life science publications is stored together with the standoff annotations of the found entities in title and abstract in a MongoDB[3] document database. The relational proper-

*Corresponding Author
[1]https://www.livivo.de/ last accessed June 2016
[2]https://www.zbmed.de/ last accessed June 2016
[3]https://www.mongodb.com/ last accessed June 2016

SEMANTICS 2016: Posters and Demos Track
September 13-14, 2016, Leipzig, Germany

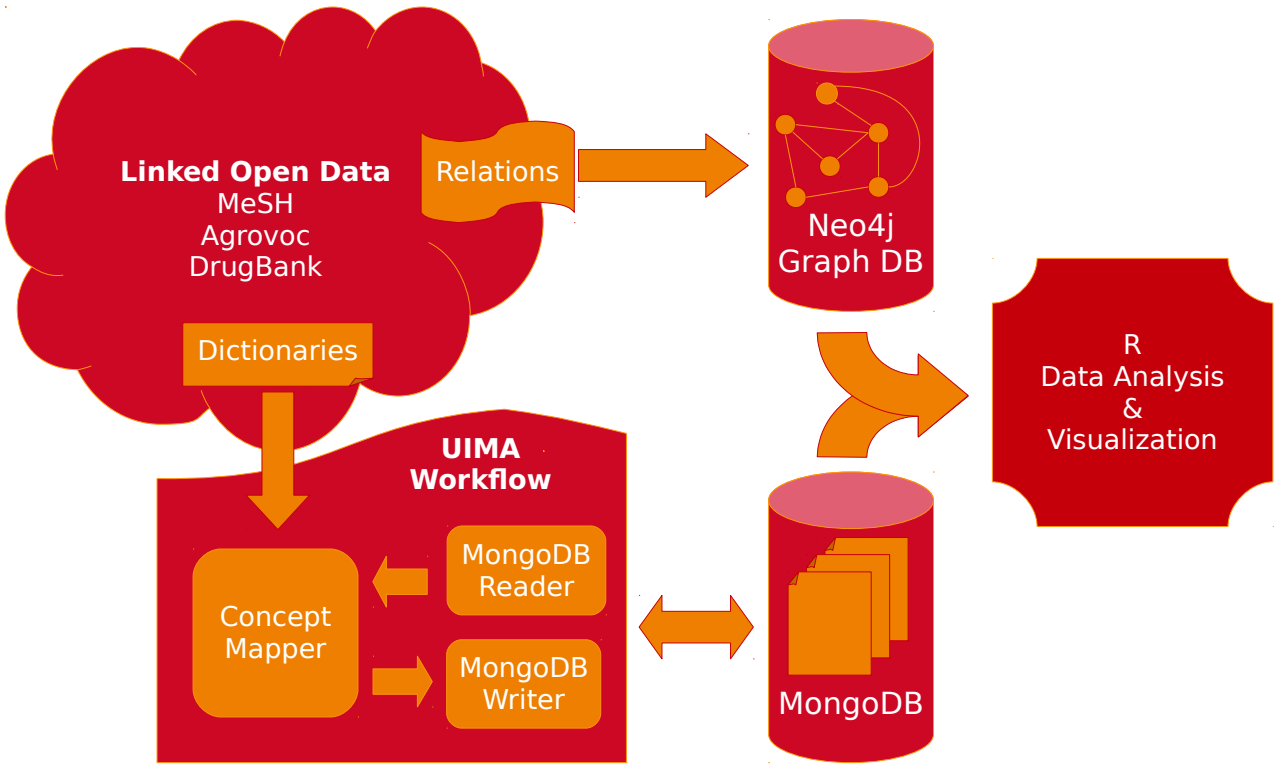

Figure 1: Data analysis workflow extracting entities from the linked data cloud for named entity recognition using UIMA and relations of the entities for the Neo4j graph database.

ties of the semantic entities are preserved in a Neo4j[4] graph database. The workflow is depicted in Figure 1. The `UIMA`-based workflow is discussed in subsection 2.1. The graph database is discussed in subsection 2.2. The data analysis and visualization using the statistical programming language R is described in section 2.4.

2.1 UIMA-based Named Entity Recognition

Entities from the linked data cloud are taken for medical, agricultural, and pharmacological terms, specifically, from the Medical Subject Headings (`MeSH`) [5] [7], the agricultural controlled vocabulary `AGROVOC`[6] [1], and `DrugBank`[7] [5]. The number of entities and synonyms are shown in Table 1.

Table 1: Number of semantic entities and their synonyms for `MeSH`, `DrugBank`, and `AGROVOC`

Dictionary	Concepts	Synonyms
`MeSH`	27,885	117,545
`DrugBank`	7,760	13,359
`Agrovoc`	32,027	N/A

Each entity carries a unique ID such as `D004827` in `MeSH` and is determined by various terms defined as synonyms such as *Epilepsy*, *Aura* or *Awakening Epilepsy*. `Agrovoc`'s focus is on multilingualism whereas the other two dictionaries from `MeSH` and `DrugBank` are primarily in English. In contrast to `MeSH` and `DrugBank`, `AGROVOC` provides a variety of translations into other languages without having synonyms in English.

2.1.1 *UIMA Workflow*

In order to automatically assign a named entity found in title or abstract, a generic `UIMA` pipeline is implemented, using the `UIMA ConceptMapper` [8]. The `UIMA` analysis process is as follows:

1. The Collection Reader `MongoDBReader` reads the unstructured text from the `LIVIVO` publications.
2. For each publication, the `Offset-Tokenizer` processes the textual information on a token-by-token basis.
3. The Analysis Engine `ConceptMapper` matches tokens as concepts. A list of annotations is compiled with information about the token. The concepts are assigned semantically with one of the linked open data databases. Furthermore the concepts are aligned with the concept ID from the respective database. Finally the offset is stored to capture the start and end position of the annotation in the metadata field.
4. The Consumer `MongoDBWriter` stores the results in the `MongoDB`.

The named entity recognition approach achieves a semantic enrichment of the unstructured and semi-structured text with information from the linked open data cloud.

[4]http://neo4j.com/ last accessed June 2016
[5]https://www.nlm.nih.gov/mesh/ last accessed June 2016
[6]http://aims.fao.org/standards/agrovoc/ last accessed June 2016
[7]http://www.drugbank.ca/ last accessed June 2016

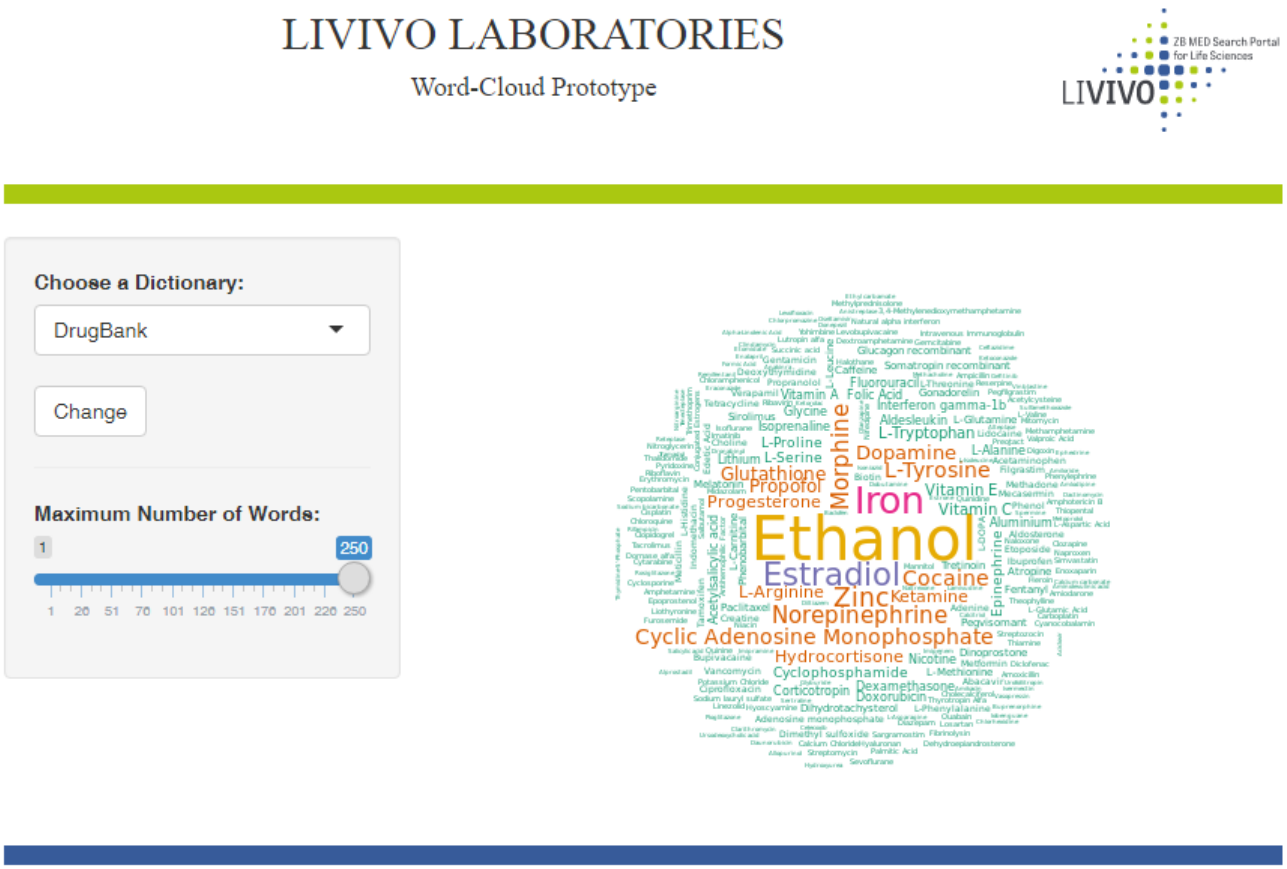

Figure 2: Prototype that visualizes concepts based on their frequency from the respective dictionaries for `MeSH`, `DrugBank`, and `Agrovoc`. The prototype is available online at `http://labs.livivo.de`

2.2 Neo4j Graph Database

The semantic entities with their relations are stored in a `Neo4j` graph database with each node representing a concept and each edge being a relationship amongst two nodes. The properties for a node comprise an internal identifier, a preferred name, the original relational identifier, and the concept identifier. In case of `DrugBank`, a node refers with its concept identifier to the DrugBank database record of the chemical compound, while the ontological identifier refers the `ATC` Classification System [3]. In case of `MeSH` and `AGROVOC`, the classification of concepts is a hierarchy with the node identifiers refering to the respective tree nodes.

2.3 MongoDB Document Database

The metadata of the 63 Million life science publications from `LIVIVO` are stored in a `MongoDB` document database. The `MongoDB` document store serves as repository for reading textual information from the documents as well as saving standoff annotations from the `UIMA`-based named entity recognition. Each document in the database comprises the metadata and the text mining information. Specifically, the metadata includes document identifiers, author information, title, abstract, and language information. The text mining annotations capture the concept identifier, the matched text, offset information as well as dictionary information.

2.4 R - Data Analysis and Visualization

Data analysis and visualization is conducted using the `R` programming language environment with packages for data integration and graphical visualizations. Both the `MongoDB` and the `Neo4j` database can be queried using their respective `R` connectors *mongolite* and *RNeo4j* for interconnecting concept properties in the documents with their relational properties. Prototypical exploration of potential visualizations for end users are developed using the *shiny* package. A `MongoDB` aggregation is created on the text mining information from each document containing standoff annotations in the `MongoDB` database. For each concept, the overall frequency is calculated as well as the total number of documents containing the concept. There are two online resources available: The prototype of the visualization and the text mining results as JSON-Linked Data:

- Prototype: `http://labs.livivo.de`
- JSON-LD: `https://datahub.io/dataset/livtdm`

A screenshot of the prototype that visualizes the word-clouds is shown in Figure 2.

3. RESULTS

After conducting a search query, search results are usually displayed as document lists. In `LIVIVO`, search results are a list of documents that are ranked according to their relevancy relative to the query terms. The ranking algorithm

in LIVIVO is TF-IDF[9]. The UIMA text and data mining workflow produces standoff annotations of semantic entities from the linked data cloud that enables the aggregation of search results according to their frequency of occurring entities in the document sets. In Table 2, the frequencies of the dictionaries for MeSH, DrugBank, and AGROVOC are shown with the total number of found concepts as well as the total number of documents containing concepts of the respective dictionary.

Table 2: Entity frequency of DrugBank, MeSH, and AGROVOC

Dictionary	# of Found Concepts	# of Documents with Concepts
MeSH	531,795,910	40,665,773
DrugBank	47,486,317	9,584,408
Agrovoc	447,766,801	39,947,272

The standoff annotations from each of the dictionaries can be used to create a word-cloud. A word-cloud is an aggregation of found entities in the documents. The word-cloud can display entities stored in the MongoDB. The most frequent entities found in publications can be displayed as well as the less used entities. It is possible to constrain the results with a variety of statistical parameters. The MongoDB database allows for visualizations and adjustments with a retrieval time on the full document corpus in the range of a few milliseconds. The frequencies for each of the word-clouds are calculated on the full corpus of LIVIVO. The frequency count is extracted from a MongoDB aggregation and piped into R using a MongoDB connector. The visualization is conducted with the visualization package *shiny* using data structures from the Text Mining packages *tm* and *wordcloud*. The most frequent entity found in both of these dictionaries is *Patients*. The most frequent entity from the *DrugBank* is *Ethanol*. As an additional visualization of aggregated search results to a word-cloud, a contextual graph can be created using the relational information from the Neo4j graph database.

4. CONCLUSIONS

The automated annotation of semantic entities in life science publications enables novel retrieval methodologies like query expansion, aggregation of search results, and concept-based ranking functions. A user can be assisted to enter the search query by auto-completing prefixes of entered search terms based on matching synonyms or concept names. Search results of a conducted query can be aggregated to co-occurring concepts in the document result set. Additionally, the co-occurring concepts can be used as filters for faceted searches. Furthermore, other ranking algorithms than TF-IDF can be implemented in LIVIVO like the concept based ranking algorithm CF-IDF[4] . Auto-completion, aggregation of search results, and concept-based ranking are currently not available in LIVIVO. It is planned to integrate these functionalities soon.

Semantic entities provide synonym resolution as well as the linkage to the major database in the life sciences because they are derived from the linked data cloud. The two different dictionaries for AGROVOC and MeSH provide the highest frequency of similar entities although the one dictionary is focused on agricultural terms while the other one is focused on medical terms. Both dictionaries also have a very high total number of found entities compared to the DrugBank dictionary that is very specific for the pharmaceutical domain of chemical compounds. Overall, the results indicate that it is important to distinguish dictionaries on a more granular level allowing a higher specificity for each of the entity classes. For example, the MeSH hierarchy provides a high specificity on the sub-root node level with 16 different categories ranging from Anatomy to Geography. In case of AGROVOC, there are 25 categories ranging from describing generic entities like *activities* to specific agricultural terms like *organisms*. Future work will have to focus on a higher specificity for the entity classes with a higher granularity of their categorizations.

5. REFERENCES

[1] C. Caracciolo, A. Stellato, A. Morshed, G. Johannsen, S. Rajbhandari, Y. Jaques, and J. Keizer. The agrovoc linked dataset. *Semantic Web*, 4(3):341–348, 2013.

[2] D. Ferrucci and A. Lally. Uima: An architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348, Sept. 2004.

[3] W. C. C. for Drug Statistics Methodology. *Guidelines for ATC classification and DDD assignment*. World Health Organization, Oslo, 2015.

[4] F. Goossen, W. IJntema, F. Frasincar, F. Hogenboom, and U. Kaymak. News personalization using the cf-idf semantic recommender. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 10:1–10:12, New York, NY, USA, 2011. ACM.

[5] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res*, 42(Database issue):D1091–D1097, Jan 2014.

[6] Z. Lu. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database (Oxford)*, 2011:baq036, 2011.

[7] F. B. Rogers. Medical subject headings. *Bull Med Libr Assoc*, 51:114–116, Jan 1963.

[8] M. Tanenblatt, A. Coden, and I. Sominsky. The conceptmapper approach to named entity recognition. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[9] S. Tuarob, L. C. Pouchard, and C. L. Giles. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, pages 239–248, New York, NY, USA, 2013. ACM.