Entity-Centric Information Access with the Human-in-the-Loop for the Biomedical Domains

Seid Muhie Yimam[†], Steffen Remus[†], Alexander Panchenko[†], Andreas Holzinger[‡], and Chris Biemann[†]

[†]Language Technology Group, Department of Informatics
Universität Hamburg, Germany
[‡]Research Unit HCI-KDD
Institute for Medical Informatics, Statistics and Documentation
Medical University Graz, Austria
{yimam,remus,panchenko,biemann}@informatik.uni-hamburg.de
a.holzinger@hci-kdd.org

Abstract

In this paper, we describe the concept of entity-centric information access for the biomedical domain. With entity recognition technologies approaching acceptable levels of accuracy, we put forward a paradigm of document browsing and searching where the entities of the domain and their relations are explicitly modeled to provide users the possibility of collecting exhaustive information on relations of interest. We describe three working prototypes along these lines: NEW/S/LEAK, which was developed for investigative journalists who need a quick overview of large leaked document collections; STORYFINDER, which is a personalized organizer for information found in web pages that allows adding entities as well as relations, and is capable of personalized information management; and adaptive annotation capabilities of WEBANNO, which is a general-purpose linguistic annotation tool. We will discuss future steps towards the adaptation of these tools to biomedical data, which is subject to a recently started project on biomedical knowledge acquisition. A key difference to other approaches is the centering around the user in a Human-in-the-Loop machine learning approach, where users define and extend categories and enable the system to improve via feedback and interaction.

1 Introduction

Recently, knowledge management as a field faced several challenges. On one hand, sophisticated technologies and standards were developed to support knowledge-based modeling, such as domain ontologies including Disease Ontology, MeSH, and Gene Ontology[1] and the Semantic Web description languages and infrastructures including RDF, OWL, SPARQL and others[2]. On the other hand, the current approaches face three major issues: (1) knowledge bottleneck: resources required for knowledge management such as domain ontologies are not available for many domains and languages; (2) the overall approach of knowledge management did not get widely spread due to the fact that it imposes a large burden on the user, such as annotation or expertise with complex tools such as Protégé[3]; (3) modeling entire domains as large as the medical domain with (English-oriented) knowledge resources does not meet requirements of users, who are mostly specializing in a certain sub-field and also need to operate in their local language.

We propose to reload this traditional heavy-weight *top-down* knowledge management approach and replace it with a much simpler and practical problem-oriented *bottom-up* approach. We choose the biomedical domain as the area of interest for our planning. Medical researchers have to process enormous amounts of literature – PubMed[4] adds about half a million papers to its index each year. Literature search and reason-

[1] http://do-wiki.nubic.northwestern.edu; http://ncbi.nlm.nih.gov/mesh; http://geneontology.org
[2] http://www.w3.org/standards/semanticweb
[3] http://protege.stanford.edu
[4] http://www.ncbi.nlm.nih.gov/pubmed

ing is demanding, because of the need to reveal and maintain many complex relationships between numerous sets of entities. In order to alleviate the efforts of biomedical research related to literature we propose a novel conception to information management based on *bottom-up* construction of a problem-oriented ontology, called entity graph (EG) in this paper. Entity graphs provide a new tool for medical researchers that (1) help to document relations between biomedical entities in a compact intuitive and interpretable form; (2) generate new relations in a semi-automatic way based on corpus analysis; (3) communicate new biomedical knowledge in a form of an easily interpretable interactive graph and (4) share knowledge and annotations amongst researchers.

2 Related Work

An early conception of a system for personal information management was Memex (Bush, 1945). The proposed design suggested that all documents of a person should be indexed to be easily accessible for consultation and for sharing with other people. Several decades later, the Web and social networks implement this vision yet only partially. According to Davenport (1994), Knowledge Management (KM) is a process of capturing, distributing, and effectively using knowledge. According to Gruber (1995), an ontology is an explicit specification of conceptualization. Studer et al. (1998) defines ontology as a formal, explicit specification of shared conceptualization. Multiple other informal and formal definitions of ontology are presented by Cimiano et al. (2014). Here "conceptualization" is a worldview, a system of conceptions and their relations.

Ontologies can be either general or domain-specific. Today's content management systems are largely accessed with facetted search, i.e. with taxonomically organized vocabularies forming a semantic facet. Users of the system must learn the vocabulary in order to assign the correct terms to newly ingested documents and to perform effective searches. The Cyc project (Lenat and Guha, 1990) was an early ontology-driven attempt to model world knowledge. Jurisica et al. (1999) presented an overview of using ontologies for information management. Later, knowledge management using ontologies was driven by the Semantic Web vision (Berners-Lee et al., 2001). This eventually led to the Linked Open Data cloud of resources, containing a comprehensive collection of interlinked ontologies. One limiting factor of widespread usage of ontologies is the heavy burden of their manual construction: all concepts, attributes and relations in ontologies are added and updated manually. Moreover, even if suitable ontologies for a target domain exist, they do not come with mechanisms to recognize their concepts in unstructured text, motivating approaches that learn ontologies from text (see Biemann (2005) and Buitelaar et al. (2005)).

Both EGs and ontologies aim at providing a shared explicit conceptualization of a certain domain. However, there are several important differences between these two resources. First, EGs are task- and/or problem-specific descriptions of a domain, while ontologies are usually designed as generic knowledge representations for a given domain. Ontologies are commonly developed as general-purpose resources that are supposed to model a certain domain without taking into account specific needs of certain application. This leads in practice to the fact that most resources should be specifically tailored to fit the need of the given task, problem or application. Along these lines, Hirst (2014) notes that the worldview captured in ontologies is based on the author of the ontology, not on the user, and the knowledge is not contextualized. We argue that this is one of the key reasons of only moderate success of ontology-based knowledge management after 15 years of development. Our approach will tackle this shortcoming: entity graphs are a knowledge representation tool that is designed to be strictly task-oriented. Such a graph would contain only concepts and relations relevant to the described problem at hand omitting any irrelevant details.

Mind maps (MMs) are visual diagrams that help to organize information about certain topics. Entity graphs have several common aspects with Mind Maps and similar knowledge management structures, such as concept maps and conceptual diagrams (Willis and Miertschin, 2006; Eppler, 2006), but are not confined to a tree structure, hence they are more apt for sharing and bring provenance in documents into the representation.

BEST is a biomedical entity search tool for knowledge discovery from biomedical literature (Lee et al., 2016). Although PubMed (the free public interface to MEDLINE, which provides access to bibliographic information in MEDLINE as

well as additional life science journals) provides a starting point to researchers, it only provides lists of relevant articles, leaving the task of extracting required information to the researchers themselves. Existing context extraction systems have limitations, such as 1) they provide outdated or incomplete results 2) the processing takes longer, and 3) most of them depend on conventional search system structures to return relevant information. BEST is developed to face the challenges of getting relevant documents from biomedical literature publications, addressing most challenges by directly returning ten relevant entities for a user's query instead of a list of documents. Our approach differs from BEST in many aspects such as 1) instead of relying on existing entity dictionaries, we use a semi-supervised entity recognition system, 2) instead of returning a pre-computed list of (indexed) results, our approach directs the researcher in pinpointing the required information with directed visual exploration, i.e. a guided search, 3) in addition to pre-defined entity types or dictionaries, our approach allows researchers to define their own entity types without the need of advanced pre-processing or text mining knowledge, i.e. adaptive annotation.

Zhang and Elhadad (2013) propose an unsupervised approach for detecting biomedical entities. Instead of hand-crafted rules or annotated dataset, this work first identifies classes of entities based on UMLS[5] semantic groups in order to collect seed terms. Next, they extract chunks in order to automatically determine named entity boundaries. Finally, they use a similarity based approach to automatically group named entities into specific semantic classes. While this approach is beneficial to identify biomedical entities, it has some drawbacks compared to our approach: 1) their approach depends on the collection of seed terms, 2) it assumes that every biomedical document is available at all times.

3 Three Technologies for Entity-Centric Information Access

While we target the biomedical domain, we will describe our previous work on other domains. The entity types might change, but the principles of the entity graph is transferable across domains.

[5]Unified Medical Language System (UMLS) is a widely used ontology of biomedical terms available at https://www.nlm.nih.gov/research/umls/.

3.1 Adaptively Annotating Entities with WEBANNO

Supervised named entity recognition (NER) systems require a substantial amount of annotated data to achieve high quality performance. We present an interactive and adaptive annotation approach. Instead of using a large sets of general purpose annotation corpora, we focus on specifically collecting high quality sets of in-domain annotations. In a case study for adaptive biomedical entity annotation, we used the automation component of WEBANNO, which is a web-based annotation tool with an online machine learning component (Yimam et al., 2014). Annotations are created in an interactive and incremental approach. The process is interactive in such a way that the tool suggests annotations that can be accepted, rejected or corrected by the annotator, whereby machine learning model gets better in time.

3.2 Case Study: Entity Annotation

We conducted an annotation task for identifying medical entities using WEBANNO automation, which is focused on B-Chronic lymphocytic leukemia (B-CLL). A medical expert selects domain related abstracts for annotation. Unlike previous approaches, the expert starts annotating texts without prior determination of the entity types. During the annotation process, important entities are identified that could help retrieving relevant documents about B-CLL. In a first step, we annotated five abstracts and use them for training to produce suggestions.

The following entity types are identified throughout the task: CELL, CONDITION, DISORDER, GENE, MOLECULE, PROTEIN, MOLECULAR PATHWAY and SUBSTANCE. We can see the following advantages of the adaptive annotation approach: 1) it makes the annotation task faster by producing correct predictions after annotating only a few number of documents, 2) the process helps the annotator to determine entity types unlike traditional approaches where the types are predefined by experts beforehand. This makes the identification of entity types more complete and robust (see details in Yimam et al., 2016a).

One of the typical relations between biomedical entities describe the cause and effect of diseases. Again, supervised machine learning approaches for automatic relation extraction requires more ef-

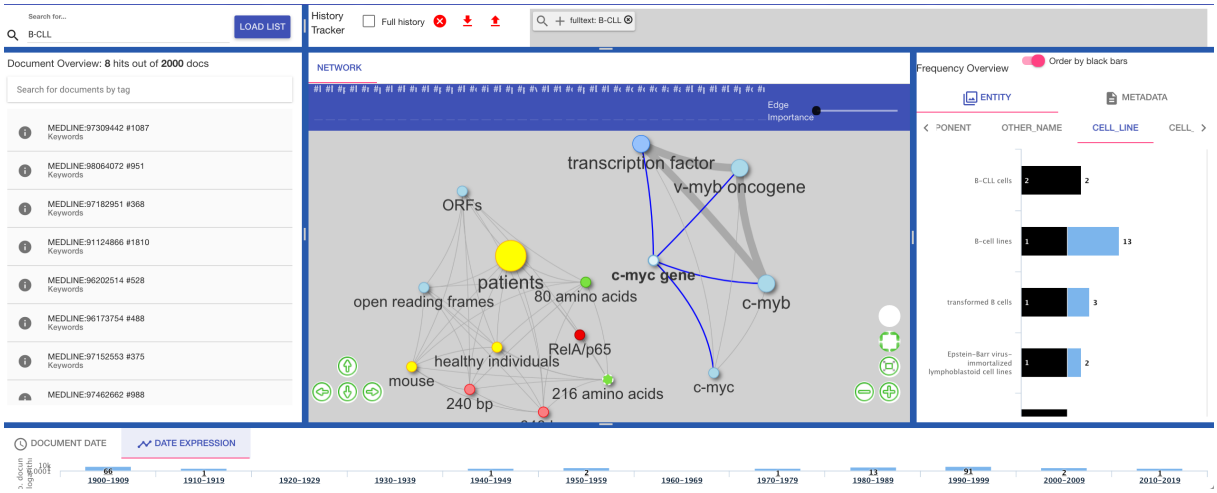


Figure 1: NEW/S/LEAK UI overview of the GENIA term annotation corpus. The example shows a B-CLL query and the graph shows involved DNA regions, "c-myc gene" is selected.

fort. For rapid annotation of relations, the relation copy annotator in WEBANNO was used, where relation suggestions are provided as soon as annotators create the first relation annotations. This functionality has the following advantages: a) experts can annotate entities as well as relation annotations at the same time, b) instances of the same entity and relation are automatically suggested for the running document as well as other unfinished documents.

3.3 Collection Insights with NEW/S/LEAK

NEW/S/LEAK is a tool designed to support investigative data journalism by exploring large sets of input documents, typically leaked documents (Yimam et al., 2016b). Named entities, such as persons, organizations, and locations, are automatically identified and ranked by importance. A global graph of entities is constructed, which is subsequently used to display high-level interactions among those entities. The tool is intended to guide investigative data journalists, by offering a rich set of possible interactions, among which are: full text search, entity merger or removal, document aggregation using meta-data, and many more.

Journalists, as targeted user group, can browse the document collection using the interactive interface (see Figure 1). It enables faceted document exploration within several views: 1) the **graph view** shows named entities and their relations, 2) the **document timeline view** shows document frequency in different epochs, 3) the **document view** is composed of the document list and a document text for reading, and 4) the **metadata views** include the search- and history views, which offer different metadata for filtering relevant or irrelevant documents.

The views are interactive, i.e. users can browse and explore the document collection on demand. The user starts with exploring entities and their connections in the graph view or by searching for entities and keywords. All interactions in the views define a filter that constrains the current document set, which in turn changes the displayed information content. User-selected entities are highlighted in the documents.

Graph view: entities and their co-occurrences The graph view shows a set of entities as nodes and their connections as links. The node size denotes the frequency of an entity, the node color denotes the entity type. The number of shown entities can be set by the user individually for each facet (entity type). The edge thickness and label denotes the size and relation of co-occurrence of the involved entities within the documents.

Document timeline The document timeline lists the number of documents in a specific epoch. Users can refine their search to see the document distribution over years, months or days.

Document view The document view shows a list of documents with their heading as selected by the currently active filters. For large document collections, the documents are loaded on demand. The document text view shows the text of the document, where the entities displayed in the graph are highlighted and underlined. The underline color corresponds to the type of entity. Selected entities in the graph are highlighted, which en-

ables a "close reading" mode to verify hypotheses formed in the so-called "distant reading" visualization (Moretti, 2007).

Metadata, search and history tracing This view is mainly used to filter documents based on different criteria such as metadata, entities, search terms/key words, etc. The history tracer helps the journalist to modify the search facets.

3.4 Personalized Knowledge Management with STORYFINDER

STORYFINDER is a toolkit that aims to keep information managed which is found and processed while browsing the web (Remus et al., 2017). The major goal is to organize a personal history of *bits of information* in form of entities and their relations rather than a history of web pages while still being able to find the source of a particular information bit in the respective web pages.

The system consists of three major components (cf. Fig. 2): 1) the Mozilla Firefox **browser plugin**, which: listens and reacts to a user's actions; initiates the analysis of a currently visited webpage on the backend server; and provides a side pane view to visualize the collected information; 2) the **server backend**, which: performs the analysis of a webpage; extracts metadata and stores the information for later access; and 3) the **interactive web page**, which: provides real-time access to the new information and is embedded in the plugin's side pane and can be accessed as a regular web page too.

In its current form, STORYFINDER is targeted for processing news texts; it automatically extracts *named entities* and draws an edge in a knowledge graph representation if two distinct entities co-occur in the same sentence (Fig. 3a).

The entities are subsequently highlighted within the current article for better visual appearance (Fig. 3b). The graph, i.e. the entities as nodes and their relations as edges are fully editable (Fig. 3c).

Due to the modular REST architecture regarding the NLP components within the backend server, every automatic component is exchangeable, e.g. in order to automatically identify medical entities such as proteins, we merely need a reliable protein tagger. In order to build such a tagger, annotated data is needed, which calls for an integration with adaptive annotation(Section 3.1).

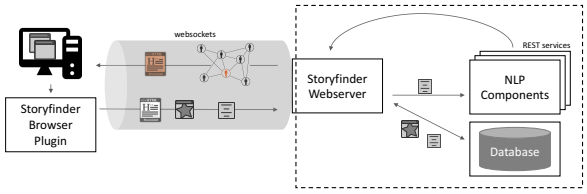
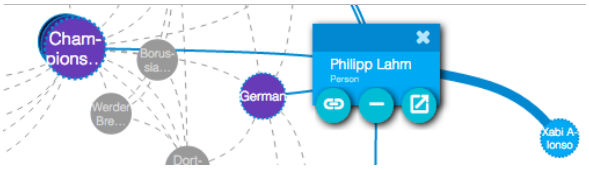
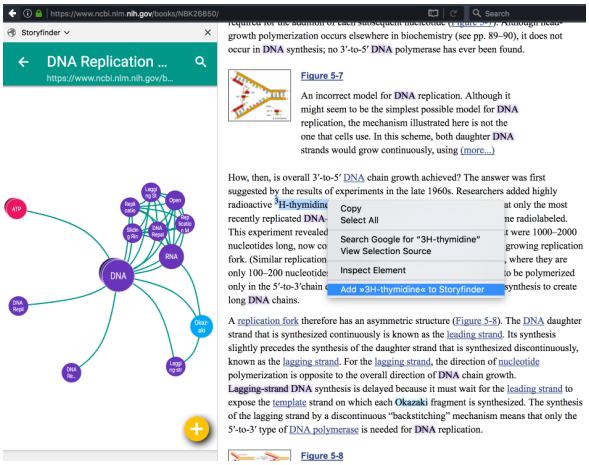


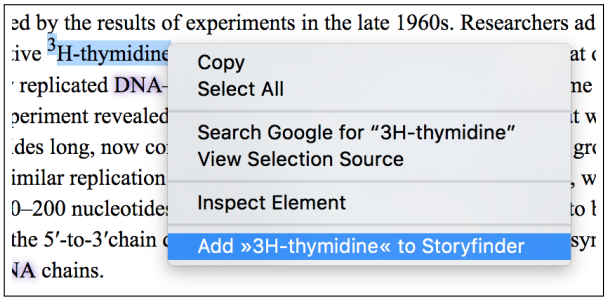
Figure 2: Schema of STORYFINDER's components: The browser plugin, the server backend, and the interactive web page.



(a) The entity 'Philipp Lahm' is selected, other nodes and edges are grayed out except direct neighboring edges and nodes. Additionally an edge is hovered (rightmost thick edge).



(b) Screenshot of the default STORYFINDER plugin view. A currently visited webpage is analyzed, and the extracted entities are highlighted in an overlay. Entities are rendered in a graph together with their relations in the STORYFINDER webpage, which is shown in a side pane of the browser.



(c) Manually adding entities of arbitrary kind can be accomplished via the plugin by right clicking any term or phrase.

Figure 3: Selected STORYFINDER screenshots.

4 Towards Information Management with Human-in-the-Loop

Within our newly started project, we will implement a prototype that uses the entity graph representation as the primary means for visualizing and

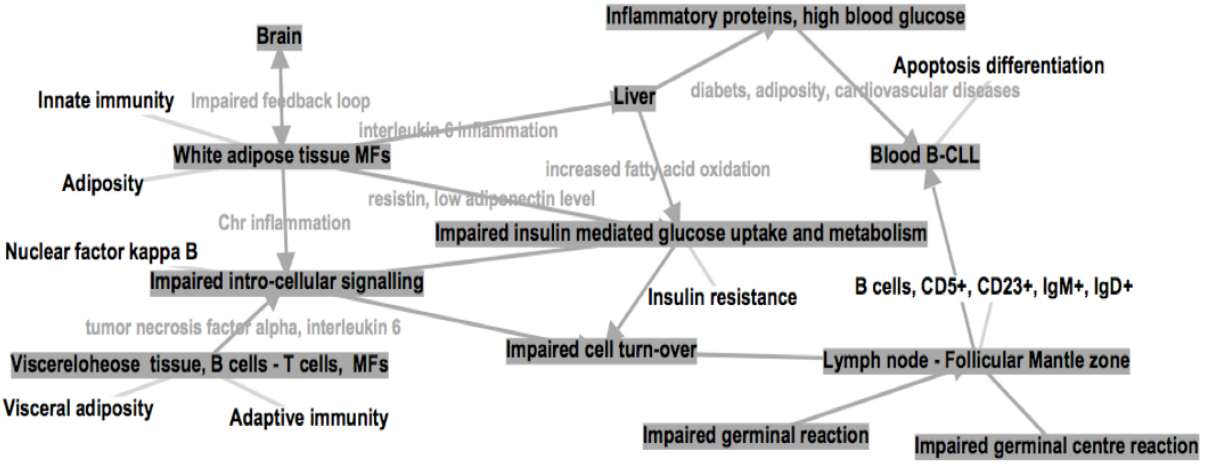


Figure 4: An entity graph summarizing the literature research on B-CLL. The key symptoms, drugs and treatments around the B-CLL are shown with their labeled relations. From labels, it becomes clear how entities relate to the topic, click on edges retrieves documents where connected concepts co-occur.

accessing biomedical research documents, integrating elements from prototypes described above. Key to the approach is to think the user in the center of the process and offer the user an adaptive ML environment (Holzinger, 2016; Holzinger et al., 2017) where manual effort in terms of annotating entities or classifying relations immediately pays of in an improved representation in the EG. To exemplify how this could look like, Figure 4 shows an example from leukemia research. Entities and their relations have been annotated and semi-automatically recognized in a personal collection of MEDLINE papers (Yimam et al., 2016a). Interacting with the network allows to find respective documents.

With these actions, the biomedical researcher can utilize the entity graph as a visually supportive notepad. Note that this goes well beyond a traditional notepad since collections of properties of entities usually do not get linked, and this also goes well beyond creativity tools such as e.g. mind maps, since it does not only displays concepts, but facilitates linking to source documents. Note further that while automatic methods aid the process, the biomedical researcher is in full control of the entity graph and can correct errors in the automatic processing in case they are relevant for the question of investigation.

Last, but not least, the individual entity graphs can be merged into a global structure by sharing among researchers. Thus in our approach, the conceptualization of a domain will be modeled from BOTTOM-UP, and not from TOP-DOWN as in the traditional knowledge management approach. Therefore, collaborative efforts of the crowd will lead to construction of a global entity graph of a domain in an incremental and problem-driven way. The global graph can be used to softly suggest edge annotations while a user constructs a new graph, making the overall process of entity graph construction backed up by a huge global entity graph, which has provenance information (i.e. who has entered information, based on which document) for mutual understanding. The global graph can also incorporate information from resources, such as MeSH, Gene and Disease Ontology. Challenges in the adaptation include a high-quality tagging of biomedical entities, preprocessing such as dependency parsing for relevant languages, the design of the user interface and a responsive online-adaptive machine learning model.

5 Conclusion

We proposed a new schema for entity-centric information extraction and -access for biomedical entities. We highlighted current drawbacks and new challenges, and presented existing tools for information extraction (WEBANNO), visualization and navigation (NEW/S/LEAK), and personalized information and knowledge management (STORYFINDER), which all together can be combined, adapted, and re-focused in order to provide a data driven, *bottom-up*, conceptualization approach. Here, the *Human-in-the-Loop* is an integral component, where not only the machine learning models for information extraction are supported and improved by users over time, the final entity graph becomes larger, cleaner, more precise and thus more usable for the users.

Acknowledgments

This research was supported by the Federal Ministry for Education and Research (Germany) under grant no. 01DS17033 and by the Volkswagen Foundation under grant no. 90 847.

References

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The Semantic Web. *Scientific American* 284(5):28–37.

Chris Biemann. 2005. Ontology learning from text - a survey of methods. *LDV Forum* (2):75–93.

Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. 2005. *Ontology learning from text : methods, evaluation and applications*, volume 123. IOS Press.

Vannevar Bush. 1945. As we may think. *The Atlantic Monthly* 176(1):101–108.

Philipp Cimiano, Christina Unger, and John McCrae. 2014. Ontology-based interpretation of natural language. *Synthesis Lectures on Human Language Technologies.* 7.2:1–178.

Thomas H. Davenport. 1994. Saving it's soul: Human-centered information management. *Harvard Business Review* 72(2):119–131.

Martin J. Eppler. 2006. A comparison between concept maps, mind maps, conceptual diagrams, and visual metaphors as complementary tools for knowledge construction and sharing. *Information Visualization* 5:202–210.

Thomas R. Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.* 43(5-6):907–928.

Graeme Hirst. 2014. *Overcoming Linguistic Barriers to the Multilingual Semantic Web*, Berlin, Heidelberg, pages 3–14.

Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3(2):119–131.

Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crisan, Camelia-M. Pintea, and Vasile Palade. 2017. A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop. *ArXiv e-prints* .

Igor Jurisica, John Mylopoulos, and Eric Yu. 1999. Using ontologies for knowledge management: An information systems perspective. In *Proceedings of the ASIS Annual Meeting*. Washington, DC, USA, pages 482–496.

Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongsoon Kim, Minji Jeon, Sangrak Lim, Donghee Choi, Sunkyu Kim, Aik-Choon Tan, and Jaewoo Kang. 2016. Best: Next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PLOS ONE* 11(10):1–16.

Douglas Lenat and Ramanathan V. Guha. 1990. *Building Large Knowledge Bases*. Addison-Wesley Pub. Co, Reading, MA.

Franco Moretti. 2007. *Graphs, maps, trees : abstract models for a literary history*. Verso, London, UK.

Steffen Remus, Manuel Kaufmann, Kathrin Ballweg, Tatiana von Landesberger, and Chris Biemann. 2017. Storyfinder: Personalized knowledge base construction and management by browsing the web. In *Proceedings of the 26th ACM International Conference on Information and Knowledge Management*. Singapore, Singapore. To appear.

Rudi Studer, V. Richard Benjamins, and Dieter Fensel. 1998. Knowledge engineering: Principles and methods. *Data Knowl. Eng.* 25(1-2):161–197.

Cheryl L. Willis and Susan L. Miertschin. 2006. Mind maps as active learning tools. *Journal of computing sciences in colleges* 21(4):266–272.

Seid Muhie Yimam, Chris Biemann, Ljiljana Majnaric, Sefket Sabanovic, and Andreas Holzinger. 2016a. An adaptive annotation approach for biomedical entity and relation recognition. *Brain Informatics* 3(3):157–168.

Seid Muhie Yimam, Richard Eckart de Castilho, Iryna Gurevych, and Chris Biemann. 2014. Automatic annotation suggestions and custom annotation layers in WebAnno. In *Proc. of ACL 2014: System Demonstrations*. Baltimore, MD, USA, pages 91–96.

Seid Muhie Yimam, Heiner Ulrich, Tatiana von Landesberger, Marcel Rosenbach, Michaela Regneri, Alexander Panchenko, Franziska Lehmann, Uli Fahrer, Chris Biemann, and Kathrin Ballweg. 2016b. new/s/leak – information extraction and visualization for investigative data journalists. In *Proceedings of ACL-2016 System Demonstrations*. Association for Computational Linguistics, Berlin, Germany, pages 163–168.

Shaodian Zhang and Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition. *J. of Biomedical Informatics* 46(6):1088–1098.