Integrated Semantic Search on Structured and Unstructured Data in the ADOnIS System

Friederike Klan, Erik Faessler,Alsayed Algergawy, Birgitta König-Ries, and Udo Hahn

Friedrich-Schiller-Universität Jena, Jena, Germany
firstname.lastname@uni-jena.de

Abstract. We introduce ADOnIS, an information system which coherently integrates two important, yet mostly disparate data sources, namely structured, tabular data, and unstructured data in terms of publications. The integration is achieved by providing the underlying background knowledge of the domains involved in terms of adequately tailored ontologies. Once the two basic data sources are semantically linked, entirely novel opportunities for cross-source information retrieval arise which we will highlight in this paper.

1 Introduction

Two mutually separated "data cultures" have emerged over the years and still persist in the field of information systems. On the one hand, the database community focuses on the *structured* representation of slices of the reality, typically in terms of relations and tables. On the other hand, the information retrieval community deals with, from a computational view, *unstructured* data, namely documents as streams of characters (and other media types, such as visual data) and tries to computationally interpret (and thus restructure) the meaning encoded in these textual data carriers. Both worlds rest on solid mathematical foundations and stable technical implementations on the basis of which huge amounts of structured and unstructured data can be managed and searched on an industrial scale. Yet, with the exception of activities aiming at the *Semantic Web* (for a survey, cf. [20]) they currently lack crossover.

This lack of integration hampers the usability of data at all levels. Consider, as a concrete example, an interdisciplinary research community such as the one established in the collaborative research center (CRC) AquaDiva, our research environment [16].[1] AquaDiva explores the role of water (Aqua) and biodiversity (Diva) for shaping the structure, properties and functions of the earth's subsurface. When a graduate student enters the CRC, she might be interested in the transport of viruses in the geological subsurface. In order to get started the student searches for an overview of the state of the art and hints what has been done on this topic in AquaDiva so far. So she searches for relevant publications in portals like PubMed or Google Scholar and poses search queries

[1] http://www.aquadiva.uni-jena.de/

to the BExIS 2 data portal, the central information system hub of the project to obtain data that have been collected already. Typically, the student will start with one query and then try to navigate results and find related entries.

Her success will strongly depend on her familiarity with the special mix of domains, skills of interacting with search engines and data repositories (including SQL/SPARQL-style query languages), her knowledge of linguistic variants and the taxonomic structures of the relevant sublanguages. For instance, queries for "virus transport subsurface", "virus transport soil", and "phages transport soil" typically return only partially overlapping result sets in PubMed or standard data management systems. This is due to simplistic string matching criteria, the incapability to account for linguistic variations of the same content (inflection variants, phrasal paraphrases, or synonyms) and the general lack of conceptual background knowledge (e.g., the taxonomic or partonomic structure of the domains' terminologies).

In our work, we aim to account for these deficiencies in a systematic way. The solution we propose is implemented in ADOnIS, the *AquaDiva Ontology-based Information System* that provides integrated and seamless access to structured data and unstructured publications by making use of a variety of semantic technologies such as ontologies and natural language processing (NLP) tools. With this, we hope to reduce the cognitive burden put on searchers while, at the same time, we intend to increase the coverage and quality of search results. In this paper, we briefly describe the methodologies underlying ADOnIS and the way users can interact with the system.

2 Related Work

Data in general and scientific data specifically can be roughly categorized into structured and unstructured data. Unstructured data has no predefined data model and is typically text-heavy. Due to its unstructured nature, it is a challenging task to extract specific and useful information [6, 10]. Retrieval algorithms for unstructured data often rely on keyword-based indexing and comparison techniques. They typically offer a search box query interface, where the searcher can input keywords of interest. Due to its simplicity, this kind of user interface, is very intuitive and easy to use. This comes at a cost. The semantics of the search query in terms of a set of input terms is not explicitly given and needs to be revealed by the information system.

On the other hand, structured data is data that is organized according to a predefined (but not necessarily explicitly known) data model, such as a table in a relational database (known data model), a document in RDF format ((partially) known data model) or a spreadsheet (unknown implicit data model). This predefined data model (if known[2]) enables search based on structured queries (e.g. SQL or SPARQL queries) with a well-known semantics. Although these kind

[2] In cases where the underlying data model is implicit (e.g. in spreadsheets), it needs to be provided by the data creator or has to be automatically extracted using machine-learning techniques. The latter can be particularly challenging, since in contrast to

of query interfaces make it easy to effectively identify and discover a piece of information and access it in concise way, they are rather complex and thus less suited to users with a non computer science background. Recent approaches have therefore started to combine and integrate *keyword-based* search approaches for unstructured data and *concept-based* approaches for structured data [3, 6, 2, 18, 19].

K-search is one of the earliest works on hybrid search that supports the retrieval of documents and knowledge [2]. The *K-search* approach aims at searching the Semantic Web as a collection of documents (unstructured data) and metadata (structured data). To achieve this goal, a hybrid strategy is proposed, where *keyword-based* and *metadata-based* search strategies are combined. *K-Search* uses two separate indexes for the hybrid search and combines the results afterwards via result intersection [10]. An ontology-based retrieval system is proposed in [6]. It adapts the classical vector space representation to be suitable for large-scale information sources. An ontology-based scheme is used to semi-automatically produce document annotations that are used for a semantic search. To cope with incomplete information in the knowledge base, the semantic search is combined with a conventional keyword-based search. Gärtner et al. [10] suggest a semantic search system (HS^3) that aims at semantically bridging the gap between structured and unstructured data. HS^3 is an automated system that augments an arbitrary knowledge base with additional information extracted from the Web. These information can then be used to build a document corpus and a combined index. This index is leveraged for a hybrid semantic search strategy that combines keyword-based and concept-based search. *TextTile* is a data visualization tool for datasets and query examination that requires a flexible analysis of structured data and unstructured text [9]. The tool includes a set of operations that can be interchangeably applied to structured as well as to unstructured textual data parts to generate useful data summaries. The tool does not make use of ontologies and semantic reasoning during the search process.

An semantic search architecture specifically designed for biodiversity data is suggested in [1]. The proposed system aims at improving the quality of the search results by exploiting ontologies and the contextual meaning of data. A mapping component links biodiversity data and concepts of a domain-specific ontology, *OntoBio*. A web interface supports end users to access data via SPARQL endpoints. In order to achieve this, the tool transforms domain ontologies, taxonomic information as well as biodiversity data into a common format. This has two disadvantages: datasets are duplicated and it becomes harder to reason on such big data. The *ELSEWEB* framework [22] aims at facilitating the integration of environmental data and providing semantic bridges between these data and species distribution models.

---

text-based documents, e.g. data tables, often reveal only scarce information that might give a hint to its meaning.

3 Overview of ADOnIS

We have implemented ADOnIS as an extension to the BExIS 2 data management platform [7][3]. In the following, we describe its two basic subsystems, namely the one dealing with already structured, tabular data (Sect. 3.1), and the one dealing with unstructured textual input on the basis of the semantic document search engine SeMedico (Sect. 3.2). The two components are supplemented by a graphical user interface that allows users to enter search terms based on which ADOnIS retrieves relevant data stored in BExIS 2 as well as publications (Sect. 4). A comprehensive view of the whole architecture of ADOnIS is provided in Fig. 1 which will be explained in the subsections to follow.

3.1 Handling Structured Data

Scientific data stored in BExIS 2 typically refer to field observations and measurements and are organized in tables. Each table and its corresponding meta information is referred to as a *dataset*. In addition to the data table containing the data values, each dataset comprises the *table schema* (name, datatype and unit of measurement for each data column) and *metadata* such as information about the data provider. Both, the actual data values and the table schema, are stored in a relational database.

To make the semantics of datasets explicit, we annotate each data table with conceptual knowledge encoded in ADOn , a domain-specific ontology expressed in Owl 2.[4] The ontology is tailored to the needs of the description of observational data from the life sciences domain. It only includes relevant classes and properties of these as TBox statements. Assertions about data values and data annotations, i.e. ABox statements, are not materialized in the ontology. Instead, we use the ontology-based data access system Ontop [5]. Based on a given ontology and a set of mappings that relate class and property symbols in the ontology to SQL views over the data in the database, Ontop provides a virtual RDF graph that can be queried using Sparql. This avoids duplication of instance data (that already reside in the relational database) and allows for sound and complete query answering in LOGSPACE under the *OWL 2 QL* entailment regime.[5] In order to retrieve datasets relevant to a certain search query, we generate a set of proper Sparql queries from the user-provided keywords, thus removing the burden from the searcher to formulate queries using a formal query language.

ADOn Ontology & Semantic Annotation. As core ontology, we use a modified version of the *Extensible Observation Ontology* (Oboe) [17] (version 1.2) that provides classes and properties for the description of field observations and measurements. Sets of related observations are organized in `oboe:Observation`

[3] `http://bexis2.uni-jena.de/`
[4] `https://www.w3.org/TR/owl-syntax`
[5] `https://www.w3.org/TR/owl-profiles/\#OWL_2_QL`

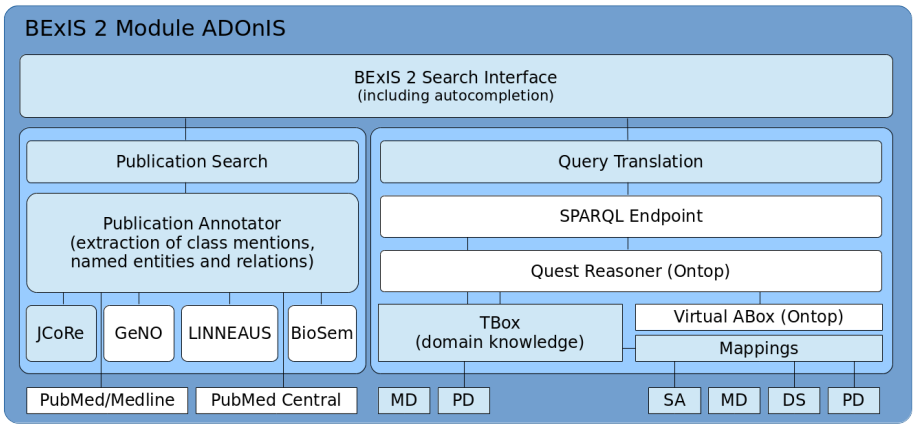

Fig. 1. System Architecture for ADOnIS

`Collection`s, which resemble the concept of a dataset in BExIS 2. Each data row in a BExIS 2 data table is modeled as one or more `oboe:Observation`s. An observation refers to an `oboe:Entity`, e.g. a *Tree*, and a set of `oboe:Measurement`s related to that entity. A measurement refers to an `oboe:Characteristic`, uses an `oboe:Standard` and results in a value. For instance, for a certain *Tree* entity, its *Circumference* (characteristic) might have been measured in meters (standard) and the measured value is 0.8. Oboe allows to indicate contextual relationships between observations, e.g. a tree might have been observed within a certain forest and this forest is located in a certain area. Modeling observations in this way enables logical inferences about entities and the relationships between them, as well as about measured characteristics of entities. In the life sciences domain, both observed entities and their characteristics are particularly important when trying to explain phenomena and thus play a key role when searching for datasets.

To cover domain-specific characteristics and entities, we reuse concepts from domain ontologies such as Obi (biomedical investigations),[6] Envo (environmental features),[7] Ncit (biomedical concepts)[8] and ChEBI (chemical entities).[9] These were selected using the Joyce tool for ontology selection and integrated into our ontology applying strict methodological criteria to guarantee non-redundancy, minimality, and optimal coverage [8]. These requirements were met by asserting subclass-relationships between concepts from a third-party ontology and either `oboe:Characteristic` or `oboe:Entity`. Since Ncit and ChEBI are huge in terms of the number of concepts they define, we used modularization techniques [8] to reuse only needed parts of these ontologies. We also

[6] `http://obi-ontology.org`
[7] `http://environmentontology.org`
[8] `https://evs.nci.nih.gov/`
[9] `https://www.ebi.ac.uk/chebi`

defined additional properties of `oboe:ObservationCollection`s, which directly relate datasets to observed entities, characteristics and standards (in contrast to OBOE, where these properties are related to individual observations). This enables efficient querying of these properties (instead of a potentially large set of observations (data rows) a much smaller number of datasets and their properties has to be inspected during search).

Each BEXIS 2 data value/data column was (manually[10]) annotated with an ontology class corresponding to the entity it refers to, an ontology class modeling the characteristic that was measured and a class referring to the measurement standard that was used. Moreover, for each dataset, we indicated contextual relationships between the observed entities. The semantic annotations are stored in a relational database.

Ontop Mappings In order to enable SPARQL queries over the conceptual view given by the ontology, we defined mappings that relate BEXIS 2 datasets, the entities and characteristics they refer to, the measured values and the dataset annotations residing in the relational database to class and property symbols in the ontology. These mappings are fixed for a given ontology and database. The subsequent mapping for example, creates a (virtual) instance for each characteristic measured in some annotated BEXIS 2 dataset. It indicates the type of this instance (some subclass of `oboe:Characteristic`) as given by the semantic annotation stored in the database table `annotation` (cf. mapping below), and relates it to dataset instances that refer to this characteristic (not depicted).

```
mappingId CHARACTERISTIC-TYPE
target :crct_{crct_id} a <{crct}> .
source SELECT DISTINCT crct, chrct_id FROM annotation
```

Query Generation Using this approach, we can pose SPARQL queries about observational data stored in BEXIS 2 on the schema level as well as on the level of individual data values. At the moment, we do not use the full potential of this solution, but rather restrict ourselves to the retrieval of BEXIS 2 datasets based on keyword queries. For that purpose, we translate the search terms into a set of SPARQL queries. For each keyword that can be mapped to the label (via string comparison) of an ontology class C that is a subclass of `oboe:Characteristic`, we create the following SPARQL query (prefixes omitted) that returns all datasets that measure C.

```
SELECT DISTINCT ?dset
WHERE {
```

[10] We are currently working on a data upload wizard which analyzes new datasets to (semi-)automatically identify semantically annotated data attributes (the type of measurement referred to in a dataset column, its datatype and unit of measurement) that are already known to and maintained by ADONIS . Such a mechanism will enable semantic annotation with little user interaction.

```
    ?dset ad:refersToCharacteristic ?char.
    ?char a <URI of C> }
```

For each keyword that can be mapped to the label of an ontology class E that is a subclass of `oboe:Entity`, this is done in a similar way, which also accounts for contextual relationships between entities. We create a SPARQL query that asks for all datasets referring to entities of type E or to some entity that appears in the context of an entity of type E.

```
SELECT DISTINCT ?dset
WHERE {
    ?dset ad:refersToEntity ?ent.
    { ?ent a <URI of D> } UNION
    { ?ent ad:hasEntityContext ?entC.
      ?entC a <URI of D> } }
```

If the label of a characteristic was entered directly before the label of an entity in the search box, we interpret this as a search for the given characteristic measured for the given entity. In case a keyword neither matches the label of an `oboe:Characteristic` nor the label of an `oboe:Entity`, we search for datasets containing data values matching the keyword. Finally, we return the union of the resulting datasets. The required information about the type of each provided keyword is delivered by an autocomplete function that provides suggestions while the user is typing words in the BExIS 2 search box. The suggestions are generated based on an index of entity and characteristic class labels defined in the underlying ontology. The keywords provided by the user as well as the keyword-related information are passed to the structured search module, which has been implemented as web service with a REST-API.

3.2 Handling Unstructured Data

Unstructured data are handled by the SEMEDICO system which receives feeds from two sources, *viz.* more than 26 million life science abstracts from MEDLINE/PUBMED[11][12] and more than 1.5 million life science full texts from PUBMED CENTRAL from the open access subset. They are stored in a POSTGRESQL database.[13]

Ontologies & Semantic Annotation. Terminological and ontological resources for the indexing of all documents come from various sources. Most notable among them is the NCBI GENE database.[14] SEMEDICO's gene recognition and normalization engine maps gene mentions in the documents to unique NCBI

[11] `https://www.ncbi.nlm.nih.gov/pubmed`
[12] `https://www.nlm.nih.gov/databases/download/pubmed_medline.html`
[13] `https://www.postgresql.org/`
[14] `https://www.ncbi.nlm.nih.gov/gene`

Gene database entries to handle gene name synonymy and ambiguity. Additionally, SeMedico integrates the Gene Ontology (Go)[15] and the Gene Regulation Ontology (Gro)[16] for the semantic description of different types of gene events.

All resources are stored in a Neo4j[17] graph database for direct access to their hierarchical structure. All terminologies, ontologies and databases are converted into a common JSON format. This format is then imported into Neo4j using a custom Neo4j server plugin.

Natural Language Processing. Before Medline and PubMed Central documents are added to SeMedico's index, they undergo an extensive linguistic analysis. The goal is to identify textual units referring to gene/protein mentions, ontology concepts, gene interaction events and factuality markers for them as expressed in the documents. To be able to recognize such higher-level semantic concepts, it is necessary to do basic linguistic analysis first like sentence and token segmentation, part-of-speech tagging and chunking.

Semantic analysis includes species tagging by the Linnaeus tagger [11], gene mention tagging and normalization using GeNo [23], gene/protein event recognition with BioSem [4] and identification of event confidence ratings following the factuality rating as described by [13]. For BioSem, we use a model trained on the BioNLP Shared Task 2011 [15] training data that includes abstracts as well as full texts. MeSH, Go and Gro concepts are tagged by a dictionary component.

All documents undergo linguistic processing employing the Uima[18] component repository JCoRe [14]. The morpho-syntactic analysis includes the resolution of acronyms [21]. This step is crucial for the interactive disambiguation feature of SeMedico. We recognize textual mentions of ontology classes via preferred names and their synonyms. When searching, also subclasses of query concepts are automatically included in the search, leveraging the ontology's subclass hierarchy. Additionally, we employ dedicated named entity recognition tools for the detection of gene / protein mentions via GeNo [23] and species via the Linneaus species tagger [11]. We also look for textually expressed relations between genes / proteins in publications. We employ BioSem [4] to extract mentions of gene / protein interactions from sentences such as

"Here we show that recombinant Pnc1 stimulates Sir2 HDAC activity."

were semantic connections between genes, proteins or, in this case, enzymes are described. Such relations have a high information value for researchers who look for interaction data on specific entities of interest. Modern relation extraction engines such as BioSem are far superior to simpler approaches which identify co-occurrences of entity within formal text units (e.g., sentences).

[15] `http://www.geneontology.org/`
[16] `https://bioportal.bioontology.org/ontologies/GRO`
[17] `https://neo4j.com/`
[18] `https://uima.apache.org/`

However, mere interaction extraction does not take into account the confidence level the authors of a publication assign to these observational data. Consider the following sentence: *"These results may suggest that mTOR-mediated autophagy inhibition may result in mesangial cell proliferation in IgAN."* While the sentence expresses some interaction between mTOR and igAN, the authors carefully use speculative words like *may* and *suggest*. Such information should be integrated into a scientific data portal to serve as an indicator how trustworthy an information item really is. We store all these annotations together with the original, raw documents in the document database.

In a last step, the analysis results required for semantic search are sent to an ElasticSearch cluster for indexing. We use a custom ElasticSearch plugin to have ElasticSearch accept a term format that allows to exactly specify index terms within the ElasticSearch index.

We model the publication search module as a web service disclosing a REST-like API. The API accepts parameters for a query string, a sorting criterion and the range of result documents that should be returned. The server then returns a JSON encoded response, including document text and bibliographic information.

4 Implementation & Preliminary Results

In this section, we introduce the GUI provided to the end user to facilitate the search process as well as preliminary evaluation results to demonstrate the effectiveness of the proposed method. To this end, we set up a running instance of the BExIS 2 system with the ADOnIS module that stores 55 real world datasets from the AquaDiva project[19]. The datasets comprise 880 data columns and 539,774 data rows in total. This results in 2,420,012 single data values. For the unstructured data search results Semedico stores more than 26M MEDLINE citations and approximately 1.5M PubMed Central full texts from the open access subset in its index.

ADOnIS comes with a graphical user interface for the semantic search (Fig. 2). It is divided into three parts: the search box (top), where the user can enter keyword queries (one or more keywords), a section displaying publications (unstructured data) relevant to the query (left) and the list of retrieved BExIS 2 datasets (structured data) (right). An exemplary search using the keywords `groundwater`, `concentration of` and `nitrate` is shown in Fig. 2. The search delivers datasets that refer to the entity groundwater or entities that have been observed in the context of groundwater and datasets where the concentration of (characteristic) nitrate (entity) was measured. On the left-hand side, relevant publications are listed.

To demonstrate the effectiveness of the search functionality of ADOnIS we compared its results to those of the original keyword-based search provided by

[19] Currently, a subset of 15 datasets including 146 data attributes has been semantically annotated.

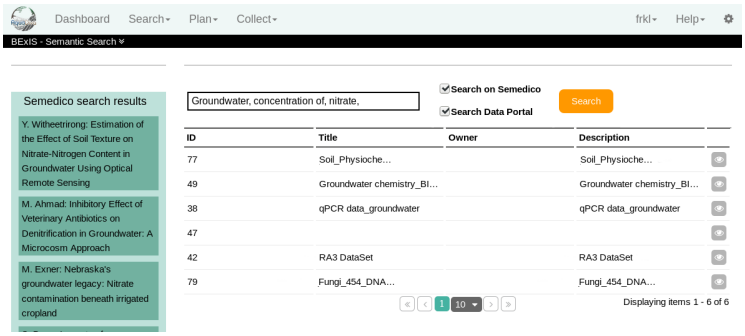

Fig. 2. Search Interface

BExIS 2 , which is powered by Apache Lucene[20] indexing both datasets and its accompanying metadata. As a preliminary evaluation, we've run the system with keyword queries relevant within the AquaDiva project. We varied the query complexity by using one or more keywords. Exemplary results are reported in Table 1. In its current version, ADOnIS returns the union of both, the results returned by the semantic search and the results retrieved by the BExIS 2 standard search. This is to avoid an empty result set in cases where the semantic search does not retrieve any (exactly fitting) datasets. As a consequence, ADOnIS can just return additional datasets that have not been found by the original BExIS 2 search.

For a single keyword, ADOnIS and BExIS 2 typically return the same results, since those keywords are often explicitly mentioned either in the datasets itself or in the metadata. However, if we consider more complex queries, ADOnIS delivers relevant results that BExIS 2 does not discover. As a next step, we will extend this preliminary evaluation. In particular, we plan to invite formal feedback from the AquaDiva researchers. This will cover both, an assessment of the relevance of the delivered search results[21] as well as an evaluation of the user interface. In addition, we will evaluate how well the search scales with an increasing number of datasets.

5 Conclusion

We introduced ADOnIS, an information system which coherently integrates two important, yet mostly disparate data sources, namely structured data from databases (or spreadsheets), on the one hand, and unstructured data in terms

[20] `https://lucene.apache.org/`

[21] Note that, even if datasets are annotated correctly, the search might deliver results that the user did not expect, since ADOnIS interprets the user's keywords in a certain way (cf. Sect. 3.1) that does not necessarily comply with the searcher's query intend. Such a mismatch would be discovered by a user study with the AquaDiva researchers.

Table 1. Search results

Keywords	**# of ADOnIS results**	**# of BExIS 2 results**
RNA	16	16
soil moisture	6	2
chemical upper aquifer	2	0
groundwater concentration of nitrate	6	0

of publications, on the other hand. The integration is achieved by providing the underlying background knowledge of the domains involved in terms of adequately tailored ontologies. Once the two basic data sources are semantically linked, entirely novel opportunities for cross-source information retrieval arise.

6 Acknowledgments

This work has been mostly funded by the *Deutsche Forschungsgemeinschaft (DFG)* as part of the CRC 1076 AquaDiva.

References

1. F. K. Amanqui, K. J. A. Serique, S. D. Cardoso, J. L. C. dos Santos, A. C. F. Albuquerque, and D. A. Moreira. Improving biodiversity data retrieval through semantic search and ontologies. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, Poland, August 11-14, 2014 - Volume II*, pages 274–281, 2014.
2. R. Bhagdev, S. Chapman, F. Ciravegna, V. Lanfranchi, and D. Petrelli. Hybrid search: Effectively combining keywords and semantic searches. In *5th European Semantic Web Conference, ESWC*, pages 554–568, 2008.
3. N. Bikakis, G. Giannopoulos, T. Dalamagas, and T. K. Sellis. Integrating keywords and semantics on document annotation and search. In *On the Move to Meaningful Internet Systems, OTM 2010 - Confederated International Conferences: CoopIS, IS, DOA and ODBASE, Hersonissos, Crete, Greece, October 25-29, 2010, Proceedings, Part II*, pages 921–938, 2010.
4. Q. C. Bui, E. M. van Mulligen, D. Campos, and J. A. Kors. A fast rule-based approach for biomedical event extraction. In *Proceedings of the BioNLP 2013 Shared Task Workshop*, pages 104–108, Sofia, Bulgaria, 2013.
5. D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, and G. Xiao. Ontop: Answering SPARQL queries over relational databases. *Semantic Web –Interoperability, Usability, Applicability*, 8(3):471–487, 2017.
6. P. Castells, M. Fernández, and D. Vallet. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Trans. Knowl. Data Eng.*, 19(2):261–272, 2007.
7. J. Chamanara and B. König-Ries. A conceptual model for data management in the field of ecology. *Ecological Informatics*, 24:261–272, 2014.

8. E. Faessler, F. Klan, A. Algergawy, B. König-Ries, and U. Hahn. Selecting and tailoring ontologies with Joyce. In *Proc. of the Intl. Conf. on Knowledge Engineering and Knowledge Management*. Springer, 2017.
9. C. Felix, A. V. Pandey, and E. Bertini. Texttile: An interactive visualization tool for seamless exploratory analysis of structured data and unstructured text. *IEEE Trans. Vis. Comput. Graph.*, 23(1):161–170, 2017.
10. M. Gärtner, A. Rauber, and H. Berger. Bridging structured and unstructured data via hybrid semantic search and interactive ontology-enhanced query formulation. *Knowl. Inf. Syst.*, 41(3):761–792, 2014.
11. M. Gerner, G. Nenadic, and C. M. Bergman. Linnaeus: a species name identification system for biomedical literature. *BMC Bioinformatics*, 11:85, 2010.
12. R. V. Guha, R. McCool, and E. Miller. Semantic search. In *Proceedings of the Twelfth International World Wide Web Conference, WWW 2003, Budapest, Hungary, May 20-24, 2003*, pages 700–709, 2003.
13. U. Hahn and C. Engelmann. Grounding epistemic modality in speakers' judgments. In D.-N. Pham and S.-B. Park, editors, *Trends in Artificial Intelligence. PRICAI 2014 –Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence. Gold Coast, Australia, 1-5 Dec, 2014*, number 8862 in Lecture Notes in Artificial Intelligence, pages 654–667. Springer, 2014.
14. U. Hahn, F. Matthies, E. Faessler, and J. Hellrich. UIMA-based JCoRe 2.0 goes GitHub and Maven Central: State-of-the-art software resource engineering and distribution of NLP pipelines. In *Proc. of the Intl. Conf. on Language Resources and Evaluation*, pages 2502–2509, Paris, 2016.
15. J. Kim, N. L. T. Nguyen, Y. Wang, J. Tsujii, T. Takagi, and A. Yonezawa. The genia event and protein coreference tasks of the bionlp shared task 2011. *BMC Bioinformatics*, 13(S-11):S1, 2012.
16. K. Küsel, K. U. Totsche, S. E. Trumbore, R. Lehmann, C. Steinhäuser, and M. Herrmann. How deep can surface signals be traced in the critical zone? merging biodiversity with biogeochemistry research in a central German Muschelkalk landscape. *frontiers in Earth Science*, 4:32, 2016.
17. J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2(3):279–296, Oct. 2007.
18. P. Peng, L. Zou, and Z. Qin. Answering top-k query combined keywords and structural queries on RDF graphs. *Inf. Syst.*, 67:19–35, 2017.
19. P. Peng, L. Zou, and D. Zhao. On the marriage of SPARQL and keywords. In *Web Technologies and Applications - 17th Asia-PacificWeb Conference, APWeb 2015, Guangzhou, China, September 18-20, 2015, Proceedings*, pages 3–16, 2015.
20. P. Ristoski and H. Paulheim. Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 36:1–22, 2016.
21. A. S. Schwartz and M. A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In *PSB 2003 – Proceedings of the Pacific Symposium on Biocomputing 2003. Kauai, Hawaii, USA, January 3-7, 2003*, pages 451–462, 2003.
22. N. Villanueva-Rosales, N. R. D. Rio, D. Pennington, and L. G. Chavira. Semantic bridges for biodiversity sciences. In *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II*, pages 310–317, 2015.
23. J. Wermter, K. Tomanek, and U. Hahn. High-performance gene name normalization with geno. *Bioinformatics*, 25(6):815–821, 2009.