Unleashing Semantics of Research Data

Florian Stegmaier[1], Christin Seifert[1], Roman Kern[2], Patrick Höfler[2],
Sebastian Bayerl[1], Michael Granitzer[1], Harald Kosch[1],
Stefanie Lindstaedt[2], Belgin Mutlu[2], Vedran Sabol[2], Kai Schlegel[1],
Stefan Zwicklbauer[1]

[1] University of Passau, Germany
[2] Know-Center, Graz, Austria

Abstract. Research depends to a large degree on the availability and quality of primary research data, i.e., data generated through experiments and evaluations. While the Web in general and Linked Data in particular provide a platform and the necessary technologies for sharing, managing and utilizing research data, an ecosystem supporting those tasks is still missing. The vision of the CODE project is the establishment of a sophisticated ecosystem for Linked Data. Here, the extraction of knowledge encapsulated in scientific research paper along with its public release as Linked Data serves as the major use case. Further, Visual Analytics approaches empower end users to analyse, integrate and organize data. During these tasks, specific Big Data issues are present.

Keywords: Linked Data, Natural Language Processing, Data Warehousing, Big Data

1 Introduction

Within the last ten years, the Web reinvented itself over and over, which led from a more or less static and silo-based Web to an open Web of data, the so called Semantic Web[3]. The main intention of the Semantic Web is to provide an open-access, machine-readable and semantic description of content mediated by ontologies. Following this, Linked Data [1] is the de-facto standard to publish and interlink distributed data sets in the Web. At its core, Linked Data defines a set of rules on how to expose data and leverages the combination of Semantic Web best practices, e.g., RDF[4] and SKOS[5].

However, the Linked Data cloud is mostly restricted to academic purposes due to unreliability of services and a lack of quality estimations of the accessible data. The vision of the CODE project[6] is to improve this situation by the creation of a web-based, commercially oriented ecosystem for the Linked Science cloud, which is the part of the Linked Data

[3] http://www.w3.org/standards/semanticweb/

[4] http://www.w3.org/RDF/

[5] http://www.w3.org/2004/02/skos/

[6] http://www.code-research.eu/

Table 1. Processable research data available in the CODE project

Type	Data Set Description
Research paper	PDF documents
Primary research data	Evaluation data of research campaigns
Retrievable data	Linked Open Data endpoints
Embedded data	Microdata, Microformat, RDFa

cloud focusing in research data. This ecosystem offers a value-creation chain to increase the interaction between all peers, e.g., data vendors or analysts. The integration of a marketplace leads on the one hand to crowd-sourced data processing and on the other hand to sustainability. By the help of provenance data central steps in the data lifecycle, e.g., creation, consumption and processing, along corresponding peers can be monitored enabling data quality estimations. Reliability in terms of retrieval will be ensured by the creation of dynamic views over certain Linked Data endpoints. The portions of data made available through those views can be queried with data warehousing functionalities serving as entry point for visual analytics applications.

The motivation behind the CODE project originated from obstacles of daily research work. When working on a specific research topic, the related work analysis is an crucial step. Unfortunately, this has to be done in a manual and time consuming way due to the following facts: First, experimental results and observations are locked in PDF documents, which are out of the box unstructured and not efficiently searchable. Second, there exist a large amount of conferences, workshops, etc. leading to an tremendous amount of published research data. Without doubt, the creation of a comprehensive overview over ongoing research activities is a cumbersome task. Moreover, these issues can lead to a complete wrong interpretation of the evolution of a research topic. Specifically for research on ad-hoc information retrieval, Armstrong et al. [2] discovered in an analysis of research papers issued within a decade, that no significant progress has been achieved. These issues could be improved by the use of the aforementioned services established by the CODE project and therefore serve as a basis for the main usage scenario.

2 Rediscovering Hidden Insights In Research

Research data is made available in various ways to the research society, e.g., stored in digital libraries or just linked to a specific website. Table 1 summarizes four data sources that are taken into account in the aforementioned usage scenario. Research papers are a valuable source of state-of-the-art knowledge. Libraries, such as of the project partner Mendeley offer terabytes of PDF documents. Apart from PDF documents raw research results are also released in a more data centric form, such as table-based data. This kind of data is mostly issued by (periodic) evaluation campaigns or computing challenges. Famous examples are the CLEF initiative[7] focusing on the promotion of research, innovation, and development of information retrieval systems, and the TPC community[8], which is performing transaction processing and database benchmarking. Both provide thousands of data points stored in Excel sheets. The two remaining data sources as depicted in Table 1 serve as sources for additional information. In addition to geographic or media related data, the Linked Open Data cloud is hosting billions of triples related to research publications as well as information from specific research areas, e.g., the biomedical domain. Embedded data (cf. Table 1) is already available in the current Web. Semantic metadata information, e.g., about authors or research interests, is embedded into websites using specific techniques and a variety of metadata vocabularies, e.g., Dublin Core[9], FOAF[10] or simply in an unstructured manner. This data can be crawled and therefore leverage provenance data.

As one can observe, there is a large amount of research data already available on the Web. The major drawback in this data landscape is the fact, that those are unconnected. Due to this fact, a comprehensive view is not possible, which leads to a loss of information. By the help of the CODE ecosystem, in particular by its data warehouse, this data gets connected and inference with respect to new knowledge is enabled.

Before considering the details of the knowledge extraction process, the correlation to the Big Data will be discussed. Dumbill states in his article [3] that the term Big Data is used as buzzword along with blurred semantics. The nature of Big Data has to be argued following the *3 Vs*: Volume, velocity and variety. As already shown, in the CODE project a high number of diverse as well as highly dynamic data sources are present each offering huge portions of data. Obviously, this situation leads to a integration problem with uncertainty in terms of data quality. Within the CODE project those issues will be tackled by crowd-sourcing.

3 Big Data Pipeline Approach

When working with Big Data, Labrinidis and Jagadish [4] argue that “we lose track of the fact that there are multiple steps to the data analysis pipeline, whether the data are big or small”. The Big Data processing pipeline proposed by CODE in terms of knowledge extraction of research data is illustrated in Figure 1.

On the left hand side of Figure [4]the data sources introduced in Section 2 serve as an input for the conceptual processing chain. The data flow (continuous arrows) as well as dependencies (dashed arrows) are also plotted in the image. The central components are *PDF analysis*, *Natural Language Processing*, *Disambiguation & Enrichment*, *Data Warehousing* and *Visual Analytics* and will be discussed in the following.

[7] http://www.clef-initiative.eu/
[8] http://www.tpc.org/
[9] http://www.dublincore.org/
[10] http://www.foaf-project.org/

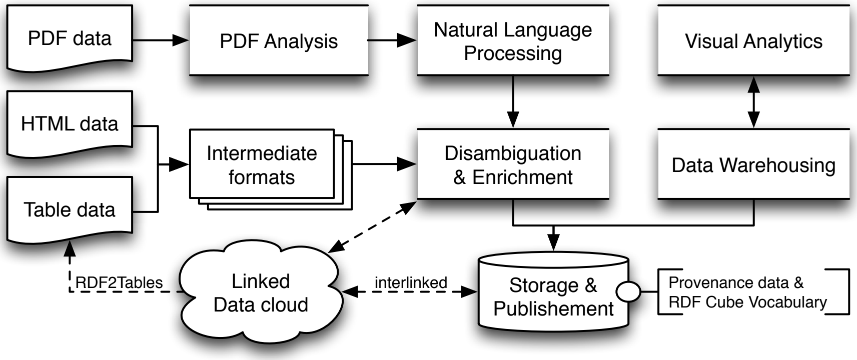

Fig. 1. Conceptual processing chain of knowledge creation and consumption

3.1 PDF Analysis

Most of the research papers are stored in the PDF format. The quality of output of the PDF analysis thereby highly influences subsequent steps in the CODE processing chain. PDF is a page description language which allows low level control of the layout, but in this process the logical structure of the text is lost. For instance, text in multiple columns is often rendered across the columns, not adhering to the natural reading order. Especially tables are challenging because there is no annotation of logical tables defined in the PDF format. Still tables are assumed to contain lot of factual quantitative information. In general the challenges for PDF analysis can be summarised as:

- Text content extraction, extracting raw textual content (ignoring images and tables).
- Metadata extraction, e.g. extracting author names, titles, journal titles for scientific publications.
- Structure annotation, annotating document structure, e.g. for generating automatic table of contents.
- Block detection, detection of logical blocks like tables, abstracts.
- Table decomposition, extraction of table data according to its logical structure.

In recent years considerable research progress has been made with regard to these challenges. Text content extraction methods are able to extract text in human-reading order [5]. Metadata extraction already quite well extracts relevant metadata from scientific papers [6, 7]. Block detection has been approached [7], but especially the extraction of complex tables is in the focus of ongoing research [8, 9].

Despite the progress in the single steps, there is no general solution which can provide all information in the quality needed within the CODE project in sufficient quality. Thus, the task is to aggregate results from recent research on PDF analysis into the CODE prototype and adapt or refine existing approaches. Further, we expect manual post-processing to be necessary for achieving certain analysis results.

3.2 Natural Language Processing

Based upon the textual representation of a research article, the contained facts should be mined. Therefore techniques from the field of natural language processing are employed. As an initial step, named entities within the text are identified. Depending on the actual domain of the articles (biomedical domain, computational science, ...) the type of named entities varies.

Domain adaptation in the CODE project is foreseen to be transformed into a crowd-sourcing task. For example, in the computer science domain, where ontologies and annotated corpora are scarce, the users of the CODE platform themselves annotate the relevant concepts. Starting with the automatic detection of named entities, the relationship between those are identified in a second step. This way the textual content is analysed and domain dependant, factual information is extracted and stored for later retrieval.

3.3 Disambiguation and Enrichment

Entity disambiguation is the task of identifying a real world entity for a given entity mentioning. In presence of a semantic knowledge base, disambiguation is the process of linking an entity to the specific entity in the knowledge base.

Within the CODE project, entity disambiguation is applied to identify and link scientific knowledge artefacts mentioned in scientific papers. Subsequently background information from the Linked Science cloud can be presented to the user while reading or writing scientific papers.

The challenges regarding entity disambiguation within the CODE project are the following: (i) variance and specificity of scientific domains: not only do scientific papers cover a wide variety of topics but each domain very in-depth; (ii) synonyms in Linked Data repositories, and (iii) evolving knowledge: topic changes in scientific papers and in Linked Data endpoints.

Disambiguation using general purpose knowledge bases (mostly Wikipedia) has been widely covered in research, e.g. [10–12]. While approaches for specific knowledge bases exist, e.g. [13] for biomedical domain, the applicability of the approaches to a combination of general and specific knowledge bases and the resulting challenges (scalability, synonyms) has to be investigated within the CODE project.

After disambiguation, the gathered information for an entity can be extended by knowledge available in the Linked Data cloud. This extra information will be validated by user feedback and then integrated into the knowledge base. This process yields to an automatic and intelligent Linked Data endpoint facing the following research tasks: (i) integration and usage of provenance data, (ii) ranking and similarity estimations of Linked Data repositories or RDF instances, and (iii) quality of service parameter (e.g., response time). This process is often termed Linked Data

Sailing. Currently, there exist frameworks to calculate similarity between Linked Data endpoints, e.g., SILK [14], and Linked Data traversal frameworks, e.g., Gremlin[11], which serves as a basis for further developments.

3.4 Storage and Publishing

The persistence layer of the CODE framework consists of a triple store, which has to offer certain abilities: (i) Linked Data compatible SPARQL endpoint and free text search capability, (ii) federated query execution, e.g., SPARQL 1.1 federated query[12], and (iii) caching strategies to ensure efficient retrieval. Those requirements are fulfilled by the Linked Media Framework [15], which has been selected for storage. For data modelling tasks, two W3C standardization efforts are in scope, which will be soon issued as official recommendations. The PROV-O[13] ontology will be used to express and interchange provenance data. Further, the W3C proposes the RDF Cube Vocabulary[14] as foundation for data cubes, which are the foundation of data warehouses. Both vocabularies will be interconnected to ensure a sophisticated retrieval process.

3.5 Data Warehousing

As already mentioned, the basis for OLAP functionalities is the data cube. The data cube model is a collection of statistical data, called observations. All observations are defined by dimensions along with measures (covering the semantics) and attributes (qualify and interpret the observation). Well-known data warehousing retrieval functionalities would last from simple aggregation functions, such as *AVG*, up to high-level *roll up* or *drill down operators*. During retrieval the following functionality has to be ensured: (i) interconnection of RDF cubes, (ii) independence of dimensions, and (iii) high-level analytical retrieval in graph structures. Current research is dealing with the integration of RDF data into single data cubes [16, 17], but do not take an interconnection / federation into scope. Within the CODE framework, algorithms of relational data warehousing systems will be evaluated with respect to their applicability to graph structures. By the help of data cube interconnections complex analytical workflows can be created.

3.6 Visual Analytics

One important aspect of the CODE project is to make data available to end users in an easy-to-use way. This data might be already Linked Data as well as semantic data extracted from scientific PDFs. The goal is to build a web-based Visual Analytics interface for users who have no prior knowledge about semantic technologies. The main challenges regarding Visual Analytics in the scope of the CODE projects are:

[11] `https://github.com/tinkerpop/gremlin/`

[12] `http://www.w3.org/TR/sparql11-federated-query/`

[13] `http://www.w3.org/TR/prov-o/`

[14] `http://www.w3.org/TR/vocab-data-cube/`

- building an easy-to-use web-based interfaces for querying, filtering and exploring semantic data,
- developing semantic descriptions of Visual Analytics components to facilitate usage with semantic data, and
- building an easy-to-use web-based interfaces for creating visual analytic dashboards.

A query wizard is envisioned, with which users can search for relevant data, filter it according to their needs, and explore and incorporate related data. Once the relevant data is selected and presented to the user in tabular form, the Visualization Wizard helps them to generate charts based on the data in order to make it easier understandable, generate new insights, and communicate those insights in a visual way. One of the tools for visualizing the data will be MeisterLabs' web-based MindMeister mind mapping platform.

4 Conclusion

In this paper the challenges of the CODE project have been outlined. Further, the connection and the relevance to Big Data topics has been argued. In the current phase of the project, prototypes for certain issues of the introduced pipeline have been developed. Within the second year of the project, those will be integrated into a single platform. Periodic evaluations will be conducted to ensure the required functionality and usability of the prototypes.

Acknowledgement The presented work was developed within the CODE project funded by the EU Seventh Framework Programme, grant agreement number 296150. The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

1. C. Bizer, T. Heath, and T. Berners-Lee, "Linked data – the story so far," *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.
2. T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel, "Improvements that don't add up: ad-hoc retrieval results since 1998.," in *Conference on Information and Knowledge Management*, pp. 601–610, 2009.
3. E. Dumbill, "What is big data? An introduction to the big data landscape." O'Reilly Strata. January 11, 2012. http://strata.oreilly.com/2012/01/what-is-big-data.html.
4. A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data.," *PVLDB*, vol. 5, no. 12, pp. 2032–2033, 2012.
5. I. Hasan, J. Parapar, and lvaro Barreiro, "Improving the extraction of text in pdfs by simulating the human reading order," *Journal of Universal Computer Science*, vol. 18, pp. 623–649, mar 2012. http://www.jucs.org/jucs_18_5/improving_the_extraction_of.

6. M. Granitzer, M. Hristakeva, R. Knight, K. Jack, and R. Kern, "A comparison of layout based bibliographic metadata extraction techniques," in *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, WIMS '12, (New York, NY, USA), pp. 19:1–19:8, ACM, 2012.
7. R. Kern, K. Jack, and M. Hristakeva, "TeamBeam - Meta-Data Extraction from Scientific Literature," *D-Lib Magazine*, vol. 18, 07 2012.
8. J. Fang, L. Gao, K. Bai, R. Qiu, X. Tao, and Z. Tang, "A table detection method for multipage pdf documents via visual seperators and tabular structures," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pp. 779 –783, sept. 2011.
9. Y. Liu, K. Bai, and L. Gao, "An efficient pre-processing method to identify logical components from pdf documents," in *Advances in Knowledge Discovery and Data Mining* (J. Huang, L. Cao, and J. Srivastava, eds.), vol. 6634 of *Lecture Notes in Computer Science*, pp. 500–511, Springer Berlin / Heidelberg, 2011.
10. S. S. Kataria, K. S. Kumar, R. R. Rastogi, P. Sen, and S. H. Sengamedu, "Entity disambiguation with hierarchical topic models," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, (New York, NY, USA), pp. 1037–1045, ACM, 2011.
11. A. Fader, S. Soderl, and O. Etzioni, "Scaling wikipediabased named entity disambiguation to arbitrary web text," in *In Proc. of WikiAI*, 2009.
12. M. Dredze, P. McNamee, D. Rao, A. Gerber, and T. Finin, "Entity disambiguation for knowledge base population," in *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, (Stroudsburg, PA, USA), pp. 277–285, Association for Computational Linguistics, 2010.
13. D. Rebholz-Schuhmann, H. Kirsch, S. Gaudan, M. Arregui, and G. Nenadic, "Annotation and disambiguation of semantic types in biomedical text: a cascaded approach to named entity recognition," in *Proceedings of the EACL Workshop on Multi-Dimensional Markup in NLP*, (Trente, Italy), 2006.
14. J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, "Discovering and maintaining links on the web of data," in *Proceedings of 8th International Semantic Web Conference*, pp. 650–665, 2009.
15. T. Kurz, S. Schaffert, and T. Bürger, "LMF – a framework for linked media," in *Proceedings of the Workshop on Multimedia on the Web collocated to i-KNOW/i-SEMANTICS*, pp. 1–4, September 2011.
16. B. Kämpgen and A. Harth, "Transforming statistical linked data for use in olap systems," in *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, (New York, NY, USA), pp. 33–40, ACM, 2011.
17. P. Zhao, X. Li, D. Xin, and J. Han, "Graph cube: on warehousing and olap multidimensional networks," in *Proceedings of the International Conference on Management of data*, pp. 853–864, 2011.