Trusted Facts: Triplifying Primary Research Data Enriched with Provenance Information*

Kai Schlegel, Sebastian Bayerl, Stefan Zwicklbauer,
Florian Stegmaier, Christin Seifert, Michael Granitzer and Harald Kosch

University of Passau, Germany
Innstrasse 41, 94032 Passau
{forename.surname}@uni-passau.de

Abstract. A crucial task in a researchers' daily work is the analysis of primary research data to estimate the evolution of certain fields or technologies, e.g. tables in publications or tabular benchmark results. Due to a lack of comparability and reliability of published primary research data, this becomes more and more time-consuming leading to contradicting facts, as has been shown for ad-hoc retrieval [1]. The CODE project [2] aims at contributing to a Linked Science Data Cloud by integrating unstructured research information with semantically represented research data. Through crowdsourcing techniques, data centric tasks like data extraction, integration and analysis in combination with sustainable data marketplace concepts will establish a *sustainable, high-impact ecosystem.*

1 Triplifying Primary Research Data

This paper demonstrates a triplification processing chain for semantic lifting of primary research data. To follow the idea of a Linked Science Data Cloud, data must be openly accessible by sophisticated retrieval possibilities to foster a dynamic interaction by the community. Figure 1 highlights essential components of this process from a conceptional point of view. If the primary research data in scope is encapsulated in unstructured data sources, a preprocessing step extracts it. For a unified view on the data, HTML table specification serves as pivot format during the semantic enrichment phase. This phase is forked in two parts: disambiguation of existent concepts and the description of the overall table structure, e.g., its dimensions or its measure type (e.g., nominal or ordinal). Both parts are crowdsourced to manage impreciseness of automatic routines applied. The initial extracted table will be enriched and annotated with semantic information using a proprietary microformat[1]. The enriched table will be finally transformed into OLAP compliant data cubes using the W3C RDF Data Cube Vocabulary[2]. Publishment is handled via a linked data endpoint following the five star deployment scheme[3].

* The presented work was developed within the CODE project funded by the EU Seventh Framework Programme, grant agreement number 296150.

[1] http://www.code-research.eu/dataextraction/microformat
[2] http://www.w3.org/TR/vocab-data-cube/
[3] http://www.5stardata.info/

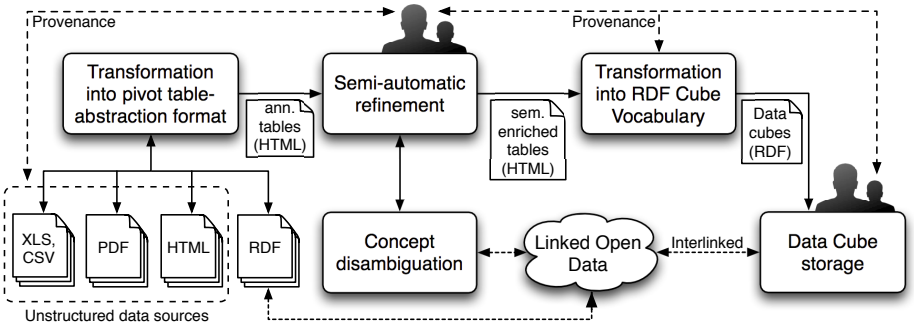

Fig. 1. Conceptual overview of data triplification chain

2 Establishment of Data Provenance Chains

A Linked Data aware publishment of primary research data is not the end of the story due to arising questions, such as: Who generated the data? Who interacted with it? The answer to these questions lead to the definition of data provenance chains enabling justification of data with respect to its impact and quality fostering trust between peers in a marketplace. As shown in Figure 1, such information is omnipresent within the triplification pipeline. To model these chains, three atomic objects are distinguished: An *entity* describes the objects whose provenance should be specified. Interaction with an entity is modeled by an *action*. Finally, the *agent* points out who is responsible for the action. Those information are then aggregated and stored by the W3C PROV ontology[4]. To add workflow specific semantics to PROV-O, the proprietary CODE Provenance Vocabulary[5] can be used.

3 Prototypic Implementation

The outlined processing chain has been implemented in the CODE Data Extraction prototype hosted at the University of Passau[6]. In the future this service will be combined with the *Testbed for Information Retrieval Algorithms (TIRA)* [3] to manage the evaluation results of CLEF challenges.

References

1. T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel, "Improvements that don't add up: ad-hoc retrieval results since 1998.," in *Proceedings of the Conference on Information and Knowledge Management*, pp. 601–610, 2009.
2. F. Stegmaier, C. Seifert, R. Kern, P. Höfler, S. Bayerl, M. Granitzer, H. Kosch, S. Lindstaedt, B. Mutlu, V. Sabol, K. Schlegel, and S. Zwicklbauer, "Unleashing semantics of research data," in *Proceedings of the 2nd Workshop on Big Data Benchmarking*, 2012.
3. T. Gollub, B. Stein, S. Burrows, and D. Hoppe, "Tira: Configuring, executing, and disseminating information retrieval experiments," *Proceedings of the 23rd International Workshop on Database and Expert Systems Applications*, pp. 151–155, 2012.

[4] http://www.w3.org/TR/prov-o/
[5] http://www.code-research.eu/ontology/code-prov-vocabulary
[6] Further details at http://www.code-research.eu/results/data-extractor.