SEDA_Lab: Towards a Laboratory for Socio-Economic Data Analysis

Johnny Ryan, Dirk Heilmann, Siegfried Handschuh

SEDA_Lab: Towards a Laboratory for Socio-Economic Data Analysis

Johnny Ryan[1], Dirk Heilmann[2] and Siegfried Handschuh[3]

[1] The Irish Times Ltd., Tara Street 24 D2, Dublin, Ireland, johnnyryan1@gmail.com
[2] HandelsBlatt Research Institute, Hohe Straße 46a 40213 Düsseldorf, Germany, d.heilmann@vhb.de
[3] Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, IDA Business Park, Lower Dangan, Galway, Ireland, siegfried.handschuh@deri.org

Abstract. In the paper we present an idea for a 'Laboratory for Socio-Economic Data Analysis' aiming to explore the use of extremely large amounts of socio-economic data from several sources of various degrees of heterogeneity. Based on these data, and on the state-of-the-art techniques in knowledge extraction and processing, we intend to deploy robust and high performance data analytics processes. Our goal is to enable the two data-bearing business partners Irish Times and Handelsblatt to use the data they have in conjunction with bigger data from external sources, to increase the value of their products and services offered and to reposition themselves in the market. We focus on two use cases that can produce tangible results in the analysis of socio-economic trends (e.g. unemployment, poverty) and socio-economic events (e.g. election tracking, bankruptcy) enabling better reporting, as well as timely decision support in crisis situations.

1 Basic Notions and the Background

The large volume of textual data available for exploitation by global user communities is growing daily. The current blogosphere hosts an average of 133 million blogs, with a daily publishing rate of 0.9 million blog posts, and with 346 million people (approx. 77% of the Internet users) reading them. The situation is analogous for many other types of resources and fields (e.g., financial data, news articles, patient records, business reports or patents).

Making the knowledge locked in these vast amounts of available texts fully accessible and interconnected is still a challenge for today's search engines and/or content management systems. The main reasons can be summarized as follows. Externalising the knowledge hidden in the individually authored texts (so that computers can efficiently process it and make it accessible) is virtually impossible when done manually. The automatic extraction of knowledge from texts has often too noisy and sparse results, which prevents their meaningful exploitation within the current knowledge representation frameworks. Hence, it is virtually impossible to integrate the results of automated knowledge extraction with legacy resources without significant (and expensive) human involvement and curation efforts.

We determine three broad types of data:

- Document corpora consisting of large amounts of structured information from metadata about the collections of documents, as well as large amounts of unstructured information represented by the textual contents of the documents themselves.
- Linked Open Data from the Web of Data - consisting of huge amounts of structured interconnected datasets, available online.
- Stream data - consisting also of online data, this time in unstructured form, captured from real-time information streams like Twitter, Facebook, blogs, news reporting agencies, stock market, financial services, etc. The data in this category is Big Data, through its volume, velocity and variety.

Based on these data, and on the state-of-the-art techniques in knowledge extraction and processing, we intend to deploy robust and high performance data analytics processes. Our goal is to enable the two data-bearing business partners Irish Times and Handelsblatt to use the data they have in conjunction with bigger data from external sources, to increase the value of their products and services offered and to reposition themselves in the market. We plan to deliver innovative and robust solutions for the representation, refinement and augmentation of the extracted knowledge, as well as for its integration. We will realise this goal through the development of novel methodologies for real-time processing, visualization and analytics over large volumes of socio-economic data (both structured and unstructured).

We focus on two use cases that can produce tangible results in the analysis of socio-economic trends (e.g. unemployment, poverty) and socio-economic events (e.g. election tracking, bankruptcy) enabling better reporting, as well as timely decision support in crisis situations.

1.1 The Adopted Approach

Our approach, in a nutshell, is to extract large amounts of knowledge from various resources and integrate it using lightweight and robust techniques with native support for data imprecision and uncertainty. These techniques will allow for storing of possibly conflicting information, resolving inconsistencies and lack of precision later on, either by means of empirical data-driven refinement, or by explicit ad-hoc interventions of users interacting via pro-active user interfaces.

More specifically, we will first deliver means for the integration and co-evolution of knowledge coming from resources of varying expressivity, noise level and relevance. In order to support that, we will build on an initial version of our novel lightweight, uncertainty-aware knowledge representation with simple similarity-based semantics easily extensible by intuitive rules. Based on these achievements, we will pursue three additional research threads:

1. Scalable multi-source knowledge extraction – Incremental knowledge extraction from textual resources and legacy terminological repositories (i.e., existing ontologies, thesauri or databases), and integration of the extracted content into continuously refined, emergent knowledge bases, which are in

turn further augmented by means of lightweight inference and used in order to boost the initial extraction results.

2. Exploiting expert communities – In addition to large-scale automatic knowledge extraction from "inanimate" resources, we will also glean much more precise content from human expert communities, their actions and contributions, integrating the gleaned content into the emergent knowledge bases in order to improve them. The integrated knowledge will serve back to the domain experts, allowing for automated detection of possibly important hidden trends and other dynamic features.
3. Pro-active user interfaces – Participation of individuals in collaborative community efforts, exploitation of the extracted knowledge and explicit individual addition or curation of the emergent knowledge will be enabled by intelligent user interfaces. These will allow for searching, visualising, browsing, but also for pro-active fetching of knowledge of interest, facilitated by profiles automatically created from the content particular users are working with. Apart from "reading" the content, the interfaces will also allow the users to "write" and thus augment that knowledge, i.e., to curate (modify) or add to the knowledge bases.

The output of the research threads described above will be implemented as an integral suite of services, accessible and interoperable. Thus we will ensure both coherence and modularity of the anticipated system, as well as easy re-use, extension for and adaptation to a large variety of use cases where textual information overload is occurs. The primary functionalities of the tool suite will be:

- search, browse and analyse knowledge extracted or inferred from textual resources of interest (either submitted manually to the system by users, or fetched automatically from external content repositories, like the information centre repositories, or the Irish Times service),
- retrieve the resources of interest based on knowledge either contained in, or pointing to them (apart from services extending the traditional key-word search),
- refine and augment the emergent knowledge bases either via one-click actions, or using intuitive machine-assisted input interfaces,
- exploit results of advanced social network analysis aimed at communities particular users are interested or directly active in.

To demonstrate and assess the applicability and usability of the results, we will be continually deploying them within two realistic complementary application scenarios in the socio-economic domain. This will be done in close cooperation with our business partners from the Irish Times and Handelsblatt, both information centres in control of huge datasets if publications in the chosen domains. The partners will continually express their case-specific requirements in an agile development process. They will also assist the technical partners and contribute with both empirical and user-based evaluation of the project deployment in the later stages.

The data that we will use in the realisation of the scenarios comes from several sources of various degrees of heterogeneity. The data is of three categories:

Document corpora - consisting of large amounts of structured information from metadata about the collections of documents, as well as large amounts of unstructured information represented by the textual contents of the documents themselves. The Irish Times and Handelsblatt provide a corpus of documents and the necessary metadata about them in the areas of the economy and the society.

This qualifies the data that we plan to use as big data as it definitely exceeds the processing capacity of conventional database systems; the data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, there is need to choose an alternative way to process it.

1.2 SEDA_Lab Overview

The general structure of both pilot use cases is depicted in the Fig. 1. There is a distinction between the producers of data and the consumers of the output knowledge. However, the same entity can be at the same time a producer and a consumer. The producers of data can be of two types, centralised and distributed. Our two business partners The Irish Times and Handelsblatt are centralised producers of data. They provide more than just the document corpora, which are an important part, but also metadata about the documents, and other already extracted annotations. The directions for the use cases are complementary. One use case looks at determining long-term trends in the information processed, for example unemployment and poverty in the social domain, or stock trading values in the economic domain. The other use case aims to examine special events, which are time-critical, like bankruptcy announcements or election result tracking. The knowledge extracted, both in regards of long-term trends and time-critical events, will be used to augment the information provided by the two information centres to their users.

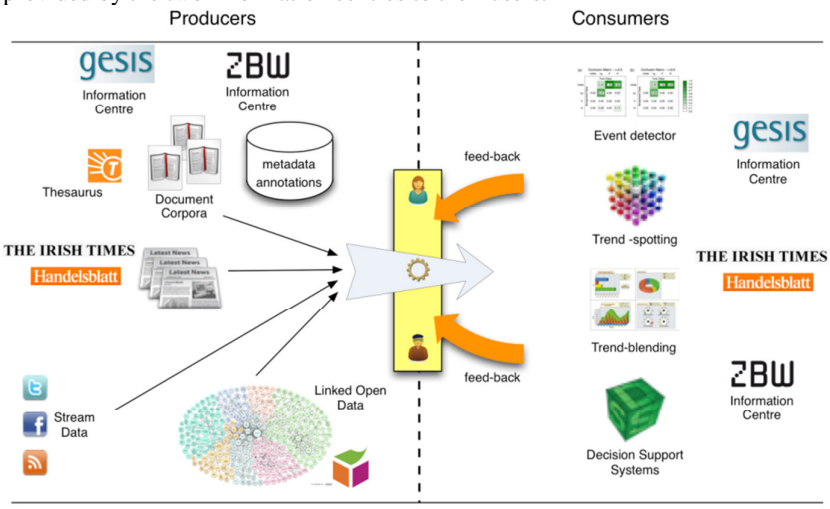

Fig. 1. Conceptual overview of the SEDA Lab interactions amongst producers and consumers

Apart from working towards the accomplishment of our research agenda, we intend to ensure maximum impact and sustainability of the results of the project, from the very beginning. The main means for achieving sustainability can be summarised as follows:

We will deliver a modular, open source and generally applicable framework for the extraction, integration, processing and delivery of knowledge coming from disparate textual or legacy resources. This will maximise the ease of deployment, re-use and extension of the researched technologies.

Building on the generally applicable framework, we will provide two specifically targeted pilot solutions. Although tailored to the requirements of two particular stakeholders, the solutions will still be transferable to many other subjects in the same areas. We will use the deployment at the major information centres, not only for mediating the respective publication knowledge to the pilot use case data consumers, but also to showcase the content delivery technologies to the customers, advertising the project's results to a potentially very large number of people.

2 Contributions to the State-of-the-Art

SEDA Lab aims at the integral advancement of machine, as well as collective and individual human intelligence, by making knowledge locked in electronic textual resources more accessible and meaningful to users who seek it. Inherent to the overall goal, the project will make the knowledge being extracted and integrated from various disparate textual sources amenable not only to exploitation, by also to open-ended intuitive augmentation or curation by individual and collaborative actors. The particular research outputs delivered in order to accomplish the goals will be implemented as a generic modular tool suite with two specific social sciences and economics pilot applications.

Automated Knowledge Extraction. Automated knowledge extraction is typically performed on two layers of text analysis. The first one deals with finding named entities and normalizing them relative to their respective database identifiers. The second layer of text analysis is concerned with the extraction of relations between named entities. Many different approaches have already been tried, among them pattern-based ones, rule-based ones, and machine learning-based ones. In the case of social sciences, an added difficulty in the task of relationship ex.

Building on various text-mining (Feldman, 2006) approaches as those outlined above, ontology learning (Maedche, 2004, Buitelaar, 2008) currently presents the most advanced method for the extraction of expressive knowledge (in terms of formal statements about entities, i.e., concepts or individuals, and their mutual relations) from natural language texts. The ontology learning approaches have recently been extended by emergent semantics principles (Maedche, 2002, Ottens, 2007) in order to construct more expressive, representative and precise ontologies (or knowledge bases, in general) in a bottom-up manner.

In addition to the text analytics, we plan to extract expressive semantic structures from traditional databases by means of data mining techniques (Dzeroski, 2001,

Hastie, 2001) wherever such datasets are available. W3C's RDB2RDF Working Group standardizes languages for mapping relational data and relational database schemas into RDF and OWL, through its two published candidate recommendations (Das, 2012, Arenas, 2012).

Emergent Knowledge Representation and Processing. Emergent knowledge is essentially dynamic and often vague or inconsistent. Approaches generalising classical formal logics have been proposed to tackle these features (Bobillo, 2008, Haase, 2005). However, the simple structure of the emergent knowledge does not allow for many non-trivial logical conclusions, rendering formal logical knowledge representations insufficient without significant human post-processing (Haase, 2005, Bechhofer, 2003). Also, even for relatively simple but large domains with many instances, logics-based querying may quickly become intractable (Hustadt, 2005). Thus, more lightweight approaches are appropriate for emergent knowledge.

Regarding Semantic Web research, numerous extensions of the basic RDF standard towards features of emergent knowledge have been investigated. Contextual features (e.g., provenance, certainty) are handled by (Schueler, 2008, Hartig, 2009). Similarity-based query post-processing with imprecision support is proposed in (Kiefer, 2007), while (Mazzieri, 2004) directly extends RDF semantics to support uncertainty (fuzzy degrees). Dynamic merging of RDF graphs is looked into by (Udrea, 2005). (Oren, 2008) researches robust approximate and scalable RDF query answering that can be used for practical exploitation of large emergent knowledge bases.

Big Data Processing. Big Data affects every sector and region of the economy and every aspect of society – providing insight into education (Dobbie, 2011), ecology (Economist, 2011) and financial risk management (Flood, 2011). It has the potential to create valuable new opportunities for individuals, communities, organisations and societies, but it also introduces new challenges (data capture, management, and analytics), and risks (privacy, security etc.). Big Data requires a new set of technologies order to transform raw data into valuable knowledge. The sheer scale of the data is an obvious challenge, however, other characteristics are equally important: the variety and velocity of the data (Russom, 2011). Variety means heterogeneity of data types, representation, and semantic interpretation. Velocity means both the rate at which data arrives and the time in which it must be acted upon.

Alternatively, (Jagadish, 2012) presents a generic pipe-line for Big Data processing, with multiple distinct phases, each important in the overall process, and each introducing new challenges: (i) acquisition, recording, (ii) extraction, cleaning, annotation, (iii) integration, aggregation, representation, (iv) analysis, modelling, (v) interpretation. Similar to this, (Fisher, 2012) describes a pipeline, focused on the architecture of the system used rather than on the process.

To tackle the challenges posed by each of the phases of the pipeline, we can make use of properties of the data, or of the characteristics of the process - the logical data independence of some databases or the inherent parallelism of some algorithms. A prominent example is Apache Mahout (mahout.apache.org) that provides implementations of a range of machine learning algorithms on Hadoop. Mahout has the drawback that it operates in batch mode and does not work effectively on stream

data. Twitter's Storm (storm-project.net) is one of a number of frameworks that has emerged for scalable analytics on real-time data.

Event and Trend Detection. There has been much research on data mining techniques for detecting interesting patterns in time series, one being event detection (a change in the values over time, which are considered quantitatively significant), but also repetitive patterns, or trends. In the statistics literature, the problem is known as "the change point detection problem" (Cherkassky, 1998) and can be treated in two ways: batch, or offline, where all the points are known and inspected together, or incremental or online, when the time points are processed one at a time (Keogh, 1997). However, in general, an event is not a single point, but a subgroup of points, which together represent the event.

Trend detection is complementary to event detection, as trends can lead to accurately predicting events even before they happen. In the project domain of socio-economic data, detecting trends like crime rates or accident hot-spots can, with the proper analysis tools, result in positive societal effects. While Twitter has been also a popular choice for trend detection (Mathioudakis, 2010, Castellanos, 2011), other equally rich sources are available to us.

3 Value Chain Thinking and the Underlying Business

Our idea for SEDA Lab builds on the principles of Value Chain Thinking (VCT) (Fearne, 2009) and reflects the natural flow of processes as these are found in the real world business environment and in a way that can support sustainable and organic growth of business ventures amongst members of the consortium. Our approach in using VCT for building the consortium and organising the workflows aims to understand stakeholders/experts attitude which after an extensive study of the field we now see that it is something not trivial at all, uncover hidden expert groups and identify hidden associations between different drivers based on the results of the above two.

As such, Value Chain Thinking is expected to help tailoring of the scenarios we deploy in the context of the project in a number of different (and with a high contextual variability) cases while still making use of the basic structure of the SEDA Lab infrastructure.

In the context of the above, SEDA Lab services can be regarded as knowledge-intensive interactions; we assume that knowledge is both the key resource for the service operations (especially as the services we deal with tend to be intangible) and a basic type of benefit/outcome received by the customer and used for the co-creation of value "in context" and "in use". In other words, we assume that the value co-creation is based on the knowledge provided by the customer (e.g. a reader of The Irish Times) to the provider (The Irish Times) and vice versa, together with the knowledge of the customer how to use in context the service in order to exploit and enhance its benefit potential.

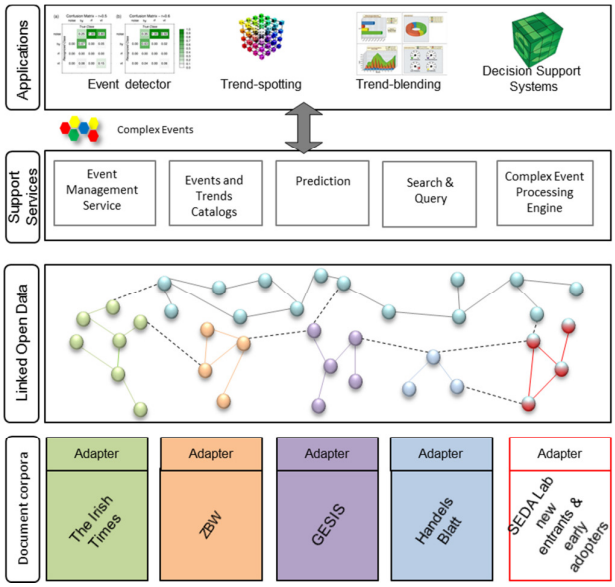

Fig. 2 Presentation of the four layers that SEDA Lab addresses. Links with other entities such as ZBW (www.zbw.eu) and GESIS (www.gesis.org) are included as well as for any new entrants in the SEDA Lab scheme.

In sum, the value of the service for the SEDA Lab customer is the result of a collaborative knowledge creation process between the customer and service providers, facilitated but not controlled or manipulated by our infrastructure. In this respect, the customer integrates all the benefits provided by the service provider, this incentivizing content owners and providers to invest from their side in building synergies amongst them and across all different steps of their data, information and knowledge value chains. In the picture above we try to visualise the layers that the adopted SEDA Lab Architecture affects:

Starting from the end user perspective, the upper layer is the one that s/he can have direct access to: these are the various Apps that a SEDA Lab user (actually: an Irish Times customer, or a Handelsblatt subscriber or registered user) may download to her/his smartphone or have access to from her/his computer at home or in office. Of course, access to these applications may happen through 3rd party applications or services through APIs. This layer is important for the project because it will be our front end to a general audience; the people will hate it or love it if the socio-economic trend-spotting, trend-blending or event detection Apps will be “cool” or have a “hot” interface.

One level down there are the Supporting services – without them no sense-making or useful App can be built. These services are built for the needs of the project by the consortium and while some of them are based on existing functionality, some others will be built for first time and will be critical for the success of the entire project. These are related to the management of socio-economic (S-E) trends and events.

The third layer deals with the Linked Open Data. Since their debut, the cloud of published data has grown considerably. Today there are 295 datasets in the form of a connected cloud, with about 31 billion RDF triples interlinked by around 504 million RDF links. Of this is only the "surfaced"/public data. More data will exists in intranets. Picking up on another example is the schema.org work. This activity from the major Web search operators will add further to the amount of machine processable.

Finally the fourth layer is the one that is provided by the Users and owners of big data in our consortium. With the trend of putting computing and networking capabilities into everyday objects and places, there are increasing numbers of data sources, producing increasing volumes of data that record what is happening in the world. Data is generated from everywhere: blogs, digital pictures, enterprise computing, feeds, social media (e.g. Twitter, Facebook), news streams and so on. This explosion of data is the result of many factors, increasing number of internet users, mobile phones, and increasing use of social networks and multimedia. Sectors such as news agencies, finance, social-economic tracking, and consumer-facing industries have also contributed significantly to the growth of Big Data.

3 Conclusions

Our idea for a laboratory for Socio-Economic Data Analysis builds on synergies with two established actors in the news publishing and news media world that will operate as big data owners and value added service providers involved in the dissemination, commercialisation and outreach activities.

SEDA Lab shall also offer a unique opportunity to both the Irish Times and Handelsblatt in terms of enabling both papers to offer a platform for comment in a variety of spheres including business, politics and public affairs, culture, the environment, health and education, to move from being newspapers of record to newspaper of reference.

References

Arenas, Marcello et al. - A Direct Mapping of Relational Data to RDF - W3C Recommendation 27 September 2012 - http://www.w3.org/TR/rdb-direct-mapping/

Bechhofer, S. et al. Tackling the ontology acquisition bottleneck: An experiment in ontology reengineering, 2003. At http://tinyurl.com/96w7ms, Apr'08.

Bobillo, F., Straccia, U. fuzzyDL: An expressive fuzzy description logic reasoner. In Proceedings of FUZZ-08, 2008.

Buitelaar, P., Cimiano, P. Ontology Learning and Population. IOS Press, 2008.

Castellanos M., Riddhiman Ghosh, Mohamed Dekhil, Perla Ruiz, Sangamesh Bellad, Umeshwar Dayal, Meichun Hsu, and Mark Schreimann. Tapping social media for sentiments in real-time. In HP TechCon 2011

Cherkassky V. and Filip Mulier. Learning from Data. Wiley-Interscience, New York, N.Y., 1998.

Das, Souripriya et al. - R2RML: RDB to RDF Mapping Language - W3C Recommendation 27 September 2012 - http://www.w3.org/TR/r2rml/

Dzeroski, S., Lavrac, N. Relational Data Mining. Springer, 2001.

Fearne A., Sustainable Food and Wine Value Chains, Adelaide Thinker in Residence, Department of the Premier and Cabinet, ISBN 978-0-9804829-9-7, 2009.

Feldman, R., Sanger, J. The Text Mining Handbook, Cambridge University Press, 2006.

Flood M., H V Jagadish, Albert Kyle, Frank Olken, and Louiqa Raschid. Using Data for Systemic Financial Risk Management. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011

Fisher, D., DeLine, R., Czerwinski, M., and Drucker, S. 2012. Interactions with big data analytics. Interactions 19, 3 (May 2012), 50-59.

Haase, P., Voelker, J. Ontology learning and reasoning - dealing with uncertainty and inconsistency. In Proceedings of the URSW2005 Workshop, pages 45–55, NOV 2005.

Hartig, O. Querying Trust in RDF Data with tSPARQL. In ESWC'09, 2009.

Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2001.

Hustadt, U., Motik, B., Sattler, U. Data complexity of reasoning in very expressive description logics. In Proc. IJCAI 2005, pages 466–471. Professional Book Center, 2005.

Jagadish, H. V. 2012. Challenges and Opportunities with Big Data,.http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf (Last accessed date: 28th June 2012)

Keogh E. and Padhraic Smyth. A probabilistic approach to fast pattern matching in time series databases. In Third International Conference on Knowledge Discovery and Data Mining, 1997.

Kiefer, C., Bernstein, A., Stocker, M. The fundamentals of isparql: A virtual triple approach for similarity-based semantic web tasks. In ISWC/ASWC, 2007.

Maedche, A. Emergent semantics for ontologies. In Emergent Semantics, IEEE Intelligent Systems. IEEE Press, 2002.

Maedche, A., Staab, S. Ontology learning. In S. Staab and R. Studer, editors, Handbook on Ontologies, chapter 9, pages 173190. Springer Verlag, 2004.

Mathioudakis M. and Nick Koudas. 2010. TwitterMonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (SIGMOD '10). ACM, New York, NY, USA, 1155-1158. DOI=10.1145/1807167.1807306 http://doi.acm.org/10.1145/1807167.1807306

Mazzieri, M. A fuzzy RDF semantics to represent trust metadata. In Proceedings of SWAP'04, 2004.

Oren, E., Gueret, C., Schlobach, S. Anytime query answering in RDF through evolutionary algorithms. In Proceedings of ISWC'08, 2008.

Ottens, K., Aussenac-Gilles, N., Gleizes, M.-P., Camps, V. Dynamic ontology co-evolution from texts: Principles and case study. In Proceedings of ESOE 2007 Workshop, pages 7083. CEUR-WS, 2007.

Russom, P.,'Big Data Analytics', TWDI Best Practices Report, Q4 2011.

Schueler, B., Sizov, S., Staab, S., Tran, D.T. Querying for meta knowledge. In Proceedings of WWW 2008. ACM, 2008.

Udrea, O., Deng, Y., Ruckhaus, E., Subrahmanian, V.S. A graph theoretical foundation for integrating RDF ontologies. In Proceedings of AAAI'05, 2005.