Joint proceedings of

Second International Workshop on Semantic Web Enterprise Adoption and Best Practice (WaSABi 2014)

&

Second International Workshop on Finance and Economics on the Semantic Web (FEOSW 2014)

Held at the 11th Extended Semantic Web Conference (ESWC 2014)

May 26th, Anissaras, Crete, Greece

WaSABi 2014: 2nd International Workshop on Semantic Web Enterprise Adoption and Best Practice

Co-located with 11th European Semantic Web Conference (ESWC 2014)

Sam Coppens, Karl Hammar, Magnus Knuth, Marco Neumann, Dominique Ritze, Miel Vander Sande

http://www.wasabi-ws.org/

1 Preface

Over the years, Semantic Web based systems, applications, and tools have shown significant improvement. Their development and deployment shows the steady maturing of semantic technologies and demonstrates their value in solving current and emerging problems. Despite the encouraging figures, the number of enterprises working on and with these technologies is dwarfed by the large number who have not yet adopted Semantic Web technologies. Current adoption is mainly restricted to methodologies provided by the research community. Although the Semantic Web acts as a candidate technology to the industry, it does not win through in current enterprise challenges. To better understand the market dynamics uptake needs to be addressed and if possible quantified.

The workshop organizer team believes that an open dialog between research and industry is beneficial and aims at a discussion in terms of best practices for enabling better market access. Consequently, WaSABi aims to guide such conversation between the scientific research community and IT practitioners with an eye towards the establishment of best practices for the development and deployment of Semantic Web based technologies. Both research and industry communities benefit from this discussion by sharing use cases, user stories, practical development issues, and design patterns.

The 2014 edition of WaSABi was in this regard a great success. The keynote speech by Marin Dimitrov positioned Semantic Web technologies on the Gartner Hype Cycle, indicating pitfalls for researchers and practitioners in this field to be aware of in the future, and suggesting approaches to help Semantic Web survive through the Trough of Disillusionment to reach the fabled Plateau of Productivity. The research papers presented touched upon a variety of topics that either prevent uptake of Semantic Web technology in industry, or could act as enablers of such uptake, including areas such as ontology quality assurance, commercially valuable information extraction, ontology design patterns, and using semantics to enable multilingual web content. Finally, the Breakout Brainstorming Session

provided a venue for discussing critical challenges for technology adoption, and developing solutions for those challenges.

We thank the authors and the program committee for their hard work in writing and reviewing papers for the workshop. We also thank our keynote speaker, Marin Dimitrov of Ontotext, for a highly relevant and interesting presentation. Finally, we thank all the workshop visitors for participating in and contributing to a successful WaSABi 2014.

June 2014

Sam Coppens
Karl Hammar
Magnus Knuth
Marco Neumann
Dominique Ritze
Miel Vander Sande

2 Organisation

2.1 Organising Committee

- Sam Coppens (IBM Research - Smarter Cities Technology Center)
- Karl Hammar (Jönköping University, Linköping University)
- Magnus Knuth (Hasso Plattner Institute - University of Potsdam)
- Marco Neumann (KONA LLC)
- Dominique Ritze (University of Mannheim)
- Miel Vander Sande (iMinds - Multimedia Lab - Ghent University)

2.2 Program Committe

- Ghislain Atemezing - Eurecom, France
- Sören Auer - University of Bonn, Fraunhofer IAIS, Germany
- Konstantin Baierer - Ex Libris, Germany
- Dan Brickley - Google, UK
- Eva Blomqvist - Linköping University, Sweden
- Andreas Blumauer - Semantic Web Company, Austria
- Frithjof Dau - SAP Research, Germany
- Johan De Smedt - Tenforce, Belgium
- Kai Eckert - University of Mannheim, Germany
- Henrik Eriksson - Linköping University, Sweden
- Daniel Garijo - Technical University of Madrid, Spain
- Peter Haase - Fluid Operations, Germany
- Corey A. Harper - New York University Libraries, USA
- Michael Hausenblas - MapR Technologies, Ireland
- Peter Mika - Yahoo! Research, Spain
- Charles McCathie Nevile - Yandex, Russia
- Heiko Paulheim - University of Mannheim, Germany
- Kurt Sandkuhl - University of Rostock, Germany
- Vladimir Tarasov - Jönköping University, Sweden
- Sebastian Tramp - AKSW – University of Leipzig, Germany
- Ruben Verborgh - iMinds – Ghent University, Belgium
- Jörg Waitelonis - Yovisto.com, Germany

3 Table of Contents

3.1 Keynote Talk

3.2 Research Papers

3.3 Breakout Session

Crossing the Chasm with Semantic Technologies

Marin Dimitrov

Ontotext AD
http://www.ontotext.com/
https://www.linkedin.com/in/marindimitrov

1 Keynote Abstract

After more than a decade of active efforts towards establishing Semantic Web, Linked Data and related standards, the verdict of whether the technology has delivered its promise and has proven itself in the enterprise is still unclear, despite the numerous existing success stories.

Every emerging technology and disruptive innovation has to overcome the challenge of "crossing the chasm" between the early adopters, who are just eager to experiment with the technology potential, and the majority of the companies, who need a proven technology that can be reliably used in mission critical scenarios and deliver quantifiable cost savings.

Succeeding with a Semantic Technology product in the enterprise is a challenging task involving both top quality research and software development practices, but most often the technology adoption challenges are not about the quality of the R&D but about successful business model generation and understanding the complexities and challenges of the technology adoption lifecycle by the enterprise.

This talk will discuss topics related to the challenge of "crossing the chasm" for a Semantic Technology product and provide examples from Ontotext's experience of successfully delivering Semantic Technology solutions to enterprises.

2 Author Bio

Marin Dimitrov is the CTO of Ontotext AD, with more than 12 years of experience in the company. His work experience includes research and development in areas related to enterprise integration systems, text mining, ontology management and Linked Data. Marin has a MSc degree in Artificial Intelligence from the University of Sofia (Bulgaria), and he is currently involved in projects related to Big Data, Cloud Computing and scalable many-core systems.

CROCUS: Cluster-based Ontology Data Cleansing

Didier Cherix[2], Ricardo Usbeck[1 2], Andreas Both[2], and Jens Lehmann[1]

[1] University of Leipzig, Germany
{usbeck,lehmann}@informatik.uni-leipzig.de
[2] R & D, Unister GmbH, Leipzig, Germany
{andreas.both,didier.cherix}@unister.de

Abstract. Over the past years, a vast number of datasets have been published based on Semantic Web standards, which provides an opportunity for creating novel industrial applications. However, industrial requirements on data quality are high while the time to market as well as the required costs for data preparation have to be kept low. Unfortunately, many Linked Data sources are error-prone which prevents their direct use in productive systems. Hence, (semi-)automatic quality assurance processes are needed as manual ontology repair procedures by domain experts are expensive and time consuming. In this article, we present CROCUS – a pipeline for cluster-based ontology data cleansing. Our system provides a semi-automatic approach for instance-level error detection in ontologies which is agnostic of the underlying Linked Data knowledge base and works at very low costs. CROCUS was evaluated on two datasets. The experiments show that we are able to detect errors with high recall.

1 Introduction

The Semantic Web movement including the Linked Open Data (LOD) cloud[1] represents a combustion point for commercial and free-to-use applications. The Linked Open Data cloud hosts over 300 publicly available knowledge bases with an extensive range of topics and DBpedia [1] as central and most important dataset. While providing a short time-to-market of large and structured datasets, Linked Data has yet not reached industrial requirements in terms of provenance, interlinking and especially data quality. In general, LOD knowledge bases comprise only few logical constraints or are not well modelled.

Industrial environments need to provide high quality data in a short amount of time. A solution might be a significant number of domain experts that are checking a given dataset and defining constraints, ensuring the demanded data quality. However, depending on the size of the given dataset the manual evaluation process by domain experts will be time consuming and expensive. Commonly, a dataset is integrated in iteration cycles repeatedly which leads to a

[1] http://lod-cloud.net/

generally good data quality. However, new or updated instances might be error-prone. Hence, the data quality of the dataset might be contaminated after a re-import.

From this scenario, we derive the requirements for our data quality evaluation process. (1) Our aim is to find singular faults, i.e., unique instance errors, conflicting with large business relevant areas of a knowledge base. (2) The data evaluation process has to be efficient. Due to the size of LOD datasets, reasoning is infeasible due to performance constraints, but graph-based statistics and clustering methods can work efficiently. (3) This process has to be agnostic of the underlying knowledge base, i.e., it should be independent of the evaluated dataset.

Often, mature ontologies, grown over years, edited by a large amount of processes and people, created by a third party provide the basis for industrial applications (e.g., DBpedia). Aiming at short time-to-market, industry needs scalable algorithms to detect errors. Furthermore, the lack of costly domain experts requires non-experts or even layman to validate the data before influencing a productive system. Resulting knowledge bases may still contain errors, however, they offer a fair trade-off in an iterative production cycle.

In this article, we present CROCUS, a cluster-based ontology data cleansing framework. CROCUS can be configured to find several types of errors in a semi-automatic way, which are afterwards validated by non-expert users called quality raters. By applying CROCUS' methodology iteratively, resulting ontology data can be safely used in industrial environments.

Our contributions are as follows: we present (1) a pipeline for semi-automatic instance-level error detection that is (2) capable of evaluating large datasets. Moreover, it is (3) an approach agnostic to the analysed class of the instance as well as the Linked Data knowledge base. Finally, (4) we provide an evaluation on a synthetic and a real-world dataset.

2 Related Work

The research field of ontology data cleansing, especially instance data can be regarded threefold: (1) development of statistical metrics to discover anomalies, (2) manual, semi-automatic and full-automatic evaluation of data quality and (3) rule- or logic-based approaches to prevent outliers in application data.

In 2013, Zaveri et al. [2] evaluate the data quality of DBpedia. This manual approach introduces a taxonomy of quality dimensions: (i) accuracy, which concerns wrong triples, data type problems and implicit relations between attributes, (ii) relevance, indicating significance of extracted information, (iii) representational consistency, measuring numerical stability and (iv) interlinking, which looks for links to external resources. Moreover, the authors present a *manual* error detection tool called *TripleCheckMate*[2] and a *semi-automatic* approach supported by the description logic learner (DL-Learner) [3,4], which generates a

[2] `http://github.com/AKSW/TripleCheckMate`

schema extension for preventing already identified errors. Those methods measured an error rate of 11.93% in DBpedia which will be a starting point for our evaluation.

A *rule-based* framework is presented by Furber et al. [5] where the authors define 9 rules of data quality. Following, the authors define an error by the number of instances not following a specific rule normalized by the overall number of relevant instances. Afterwards, the framework is able to generate statistics on which rules have been applied to the data. Several *semi-automatic* processes, e.g., [6,7], have been developed to detect errors in instance data of ontologies. Bohm et al. [6] profiled LOD knowledge bases, i.e., *statistical* metadata is generated to discover outliers. Therefore, the authors clustered the ontology to ensure partitions contain only semantically correlated data and are able to detect outliers. Hogan et al. [7] only identified errors in RDF data without evaluating the data properties itself.

In 2013, Kontokostas et al. [8] present an *automatic* methodology to assess data quality via a SPARQL-endpoint[3]. The authors define 14 basic graph patterns (BGP) to detect diverse error types. Each pattern leads to the construction of several cases with meta variables bound to specific instances of resources and literals, e.g., constructing a SPARQL query testing that a person is born before the person dies. This approach is not able to work iteratively to refine its result and is thus not usable in circular developement processes.

A first classification of quality dimensions is presented by Wang et al. [9] with respect to their importance to the user. This study reveals a classification of data quality metrics in four categories. Recently, Zaveri et al. [10] presents a systematic literature review on different methodologies for data quality assessment. The authors chose 21 articles, extracted 26 quality dimensions and categorized them according to [9]. The resulting overview shows which error types exist and whether they are repairable manually, semi-automatic or fully automatic. The presented measures were used to classify CROCUS.

To the best of our knowledge, our tool is the first tool tackling error accuracy (intrinsic data quality), completeness (contextual data quality) and consistency (data modelling) at once in a semi-automatic manner reaching high f1-measure on real-world data.

3 Method

First, we need a standardized extraction of target data to be agnostic of the underlying knowledge base. SPARQL [11] is a W3C standard to query instance data from Linked Data knowledge bases. The `DESCRIBE` query command is a way to retrieve descriptive data of certain instances. However, this query command depends on the knowledge base vendor and its configuration. To circumvent knowledge base dependence, we use *Concise Bounded Descriptions* (CBD) [12]. Given a resource r and a certain description depth d the CBD works as follows:

[3] `http://www.w3.org/TR/rdf-sparql-query/`

(1) extract all triples with r as subject and (2) resolve all blank nodes retrieved so far, i.e., for each blank node add every triple containing a blank node with the same identifier as a subject to the description. Finally, CBD repeats these steps d times. CBD configured with $d = 1$ retrieves only triples with r as subject although triples with r as object could contain useful information. Therefore, a rule is added to CBD, i.e., (3) extract all triples with r as object, which is called *Symmetric Concise Bounded Description* (SCDB) [12].

Second, CROCUS needs to calculate a numeric representation of an instance to facilitate further clustering steps. Metrics are split into three categories:

(1) The simplest metric counts each property (*count*). For example, this metric can be used if a person is expected to have only one telephone number.

(2) For each instance, the range of the resource at a certain property is counted (*range count*). In general, an undergraduate student should take undergraduate courses. If there is an undergraduate student taking courses with another type (e.g., graduate courses), this metric is able to detect it.

(3) The most general metric transforms each instance into a numeric vector and normalizes it (*numeric*). Since instances created by the SCDB consist of properties with multiple ranges, CROCUS defines the following metrics: (a) numeric properties are taken as is, (b) properties based on strings are converted to a metric by using string length although more sophisticated measures could be used (e.g., n-gram similarities) and (c) object properties are discarded for this metric.

As a third step, we apply the *density-based spatial clustering of applications with noise* (DBSCAN) algorithm [13] since it is an efficient algorithm and the order of instances has no influence on the clustering result. DBSCAN clusters instances based on the size of a cluster and the distance between those instances. Thus, DBSCAN has two parameters: ϵ, the distance between two instances, here calculated by the metrics above and $MinPts$, the minimum number of instances needed to form a cluster. If a cluster has less than $MinPts$ instances, they are regarded as outliers. We report the quality of CROCUS for different values of $MinPts$ in Section 4.

Finally, identified outliers are extracted and given to human quality judges. Based on the revised set of outliers, the algorithm can be adjusted and constraints can be added to the Linked Data knowledge base to prevent repeating discovered errors.

4 Evaluation

LUBM benchmark. First, we used the LUBM benchmark [14] to create a perfectly modelled dataset. This benchmark allows to generate arbitrary knowledge bases themed as university ontology. Our dataset consists of exactly one university and can be downloaded from our project homepage[4].

The LUBM benchmark generates random but error free data. Thus, we add different errors and error types manually for evaluation purposes:

[4] `https://github.com/AKSW/CROCUS`

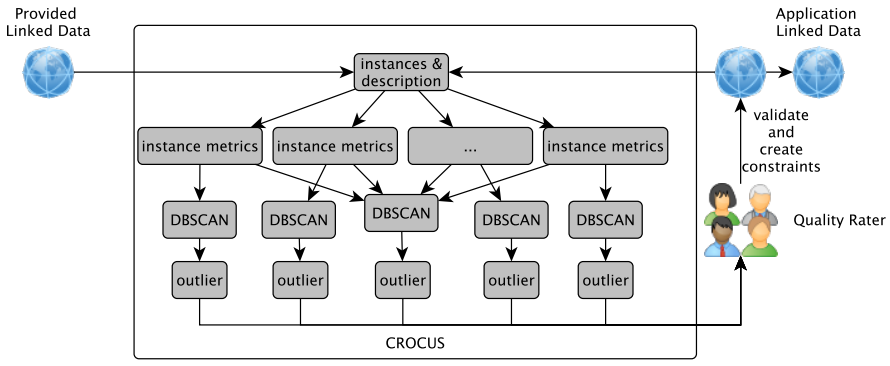


Fig. 1: Overview of CROCUS.

- *completeness of properties (count)* has been tested with CROCUS by adding a second phone number to 20 of 1874 graduate students in the dataset. The edited instances are denoted as I_{count}.
- *semantic correctness of properties (range count)* has been evaluated by adding for non-graduate students (`Course`) to 20 graduate students ($I_{rangecount}$).
- *numeric correctness of properties (numeric)* was injected by defining that a graduate student has to be younger than a certain age. To test this, 20 graduate students ($I_{numeric}$) age was replaced with a value bigger than the arbitrary maximum age of any other graduate.

For each set of instances holds: $|I_{count}| = |I_{rangecount}| = |I_{numeric}| = 20$ and additionally $|I_{count} \cap I_{rangecount} \cap I_{numeric}| = 3$. The second equation overcomes a biased evaluation and introduces some realistic noise into the dataset. One of those 3 instances is shown in the listing below:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ns2: <http://example.org/#> .
@prefix ns3: <http://www.Department6.University0.edu/> .

ns3:GraduateStudent75 a ns2:GraduateStudent ;
  ns2:name "GraduateStudent75" ;
  ns2:undergraduateDegreeFrom <http://www.University467.edu> ;
  ns2:emailAddress "GraduateStudent75@Department6.University0.edu" ;
  ns2:telephone "yyyy-yyyy-yyyy" , "xxx-xxx-xxxx" ;
  ns2:memberOf <http://www.Department6.University0.edu> ;
  ns2:age "63" ;
  ns2:takesCourse ns3:GraduateCourse21 , ns3:Course39 , ns3:
      GraduateCourse26 ;
  ns2:advisor ns3:AssociateProfessor8 .
```

Listing 1.1: Example of an instance with manually added errors (*in red*).

DBpedia - German universities benchmark. Second, we used a subset of the English DBpedia 3.8 to extract all German universities. The following SPARQL query (Listing 1.2) presents already the difficulty to find a complete list of universities using DBpedia.

```
SELECT DISTINCT ?instance
    WHERE {
        {   ?instance a dbo:University.
            ?instance dbo:country dbpedia:Germany.
            ?instance foaf:homepage ?h.
        } UNION {
            ?instance a dbo:University.
            ?instance dbp::country dbpedia:Germany.
            ?instance foaf:homepage ?h.
        } UNION {
            ?instance a dbo:University.
            ?instance dbp::country "Germany"@en.
            ?instance foaf:homepage ?h.
        }}
```

Listing 1.2: SPARQL query to extract all German universities.

After applying CROCUS to the 208 universities and validating detected instances manually, we found 39 incorrect instances. This list of incorrect instances, i.e., CBD of URIs, as well as the overall dataset can be found on our project homepage. For our evaluation, we used only properties existing in at least 50% of the instances to reduce the exponential parameter space. Apart from an increased performance of CROCUS we did not find any effective drawbacks on our results.

Results. To evaluate the performance of CROCUS, we used each error type individually on the adjusted LUBM benchmark datasets as well as a combination of all error types on LUBM[5] and the real-world DBpedia subset.

	LUBM								
	count			***range count***			***numeric***		
MinPts	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
2	—	—	—	—	—	—	—	—	—
4	—	—	—	0.49	1.00	0.33	—	—	—
8	—	—	—	0.67	1.00	0.5	—	—	—
10	0.52	1.00	0.35	1.00	1.00	1.00	—	—	—
20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 1: Results of the LUBM benchmark for all three error types.

Table 1 shows the f1-measure (F1), precision (P) and recall (R) for each error type. For some values of $MinPts$ it is infeasible to calculate cluster since DBSCAN generates only clusters but is unable to detect outlier. CROCUS is able to detect the outliers with a 1.00 f1-measure as soon as the correct size of $MinPts$ is found.

[5] The datasets can also be found on our project homepage.

Table 2 presents the results for the combined error types as well as for the German universities DBpedia subset. Combining different error types yielding a more realistic scenario influences the recall which results in a lower f1-measure than on each individual error type. Finding the optimal $MinPts$ can efficiently be done by iterating between $[2, \ldots, |I|]$. However, CROCUS achieves a high recall on the real-world data from DBpedia. Reaching a f1-measure of 0.84 for LUBM and 0.91 for DBpedia highlights CROCUS detection abilities.

	LUBM			DBpedia		
MinPts	**F1**	**P**	**R**	**F1**	**P**	**R**
2	0.12	1.00	0.09	0.04	0.25	0.02
4	0.58	1.00	0.41	0.04	0.25	0.02
8	0.84	1.00	0.72	0.04	0.25	0.02
10	0.84	1.00	0.72	0.01	0.25	0.01
20	0.84	1.00	0.72	0.17	0.44	0.10
30	0.84	1.00	0.72	0.91	0.86	0.97
50	0.84	1.00	0.72	0.85	0.80	0.97
100	0.84	1.00	0.72	0.82	0.72	0.97

Table 2: Evaluation of CROCUS against a synthetic and a real-world dataset using all metrics combined.

Property	**Errors**
`dbp:staff`, `dbp:established`, `dbp:internationalStudents`	Values are typed as `xsd:string` although they contain numeric types like integer or double.
`dbo:country`, `dbp:country`	`dbp:country` "Germany"@en collides with `dbo:Germany`

Table 3: Different error types discovered by quality raters using the German universities DBpedia subset.

In general, CROCUS generated many candidates which were then manually validated by human quality raters, who discovered a variety of errors. Table 3 lists the identified reasons of errors from the German universities DBpedia subset detected as outlier. As mentioned before, some universities do not have a property `dbo:country`. However, we found a new type of error. Some literals are of type `xsd:string` although they represent a numeric value. Lists of wrong instances can also be found on our project homepage.

Overall, CROCUS has been shown to be able to detect outliers in synthetic and real-world data and is able to work with different knowledge bases.

5 Conclusion

We presented CROCUS, a novel architecture for cluster-based, iterative ontology data cleansing, agnostic of the underlying knowledge base. With this approach we aim at the iterative integration of data into a productive environment which is a typical task of industrial software life cycles.

The experiments showed the applicability of our approach on a synthetic and, more importantly, a real-world Linked Data set. Finally, CROCUS has already been successfully used on a travel domain-specific productive environment comprising more than 630.000 instances (the dataset cannot be published due to its license).

In the future, we aim at a more extensive evaluation on domain specific knowledge bases. Furthermore, CROCUS will be extended towards a pipeline comprising a change management, an open API and semantic versioning of the underlying data. Additionally, a guided constraint derivation for laymen will be added.

Acknowledgments This work has been partly supported by the ESF and the Free State of Saxony and by grants from the European Union's 7th Framework Programme provided for the project GeoKnow (GA no. 318159). Sincere thanks to Christiane Lemke.

References

1. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. SWJ (2014)
2. Zaveri, A., Kontokostas, D., Sherif, M.A., Bühmann, L., Morsey, M., Auer, S., Lehmann, J.: User-driven quality evaluation of dbpedia. In Sabou, M., Blomqvist, E., Noia, T.D., Sack, H., Pellegrini, T., eds.: I-SEMANTICS, ACM (2013) 97–104
3. Lehmann, J.: DL-learner: Learning concepts in description logics. Journal of Machine Learning Research **10** (2009) 2639–2642
4. Buhmann, L., Lehmann, J.: Pattern based knowledge base enrichment. In: 12th ISWC, 21-25 October 2013, Sydney, Australia. (2013)
5. Fürber, C., Hepp, M.: Swiqa - a semantic web information quality assessment framework. In Tuunainen, V.K., Rossi, M., Nandhakumar, J., eds.: ECIS. (2011)
6. Böhm, C., Naumann, F., Abedjan, Z., Fenz, D., Grutze, T., Hefenbrock, D., Pohl, M., Sonnabend, D.: Profiling linked open data with ProLOD. Data Engineering Workshops ICDEW 2010 IEEE 26th International Conference on (2010) 175–178
7. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In Bizer, C., Heath, T., Berners-Lee, T., Hausenblas, M., eds.: LDOW. Volume 628 of CEUR Workshop Proceedings., CEUR-WS.org (2010)
8. Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A.J.: Test-driven evaluation of linked data quality. In: Proceedings of the 23rd international conference on World Wide Web. (2014) to appear.
9. Wang, R.Y., Strong, D.M.: Beyond accuracy. what data quality means to data consumers. Journal of Management Information Systems (4) 5–33
10. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S., Hitzler, P.: Quality assessment methodologies for linked open data. Submitted to SWJ (2013)
11. Quilitz, B., Leser, U.: Querying distributed rdf data sources with sparql. In Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M., eds.: The Semantic Web: Research and Applications. Volume 5021 of Lecture Computer Science. Springer Berlin Heidelberg (2008) 524–538
12. Stickler, P.: Cbd-concise bounded description. W3C Member Submission **3** (2005)
13. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD. Volume 96. (1996) 226–231
14. Guo, Y., Pan, Z., Heflin, J.: LUBM: A benchmark for OWL knowledge base systems. Web Semantics: Science, Services and Agents on the World Wide Web **3**(2–3) (2005) 158 – 182

IRIS: A Protégé Plug-in to Extract and Serialize Product Attribute Name-Value Pairs

Tuğba Özacar

Department of Computer Engineering, Celal Bayar University
Muradiye, 45140, Manisa, Turkey
tugba.ozacar@cbu.edu.tr

Abstract. This article introduces IRIS wrapper, which is developed as a Protégé plug-in, to solve an increasingly important problem: extracting information from the product descriptions provided by online sources and structuring this information so that is sharable among business entities, software agents and search engines. Extracted product information is presented in a GoodRelations-compliant ontology. IRIS also automatically marks up your products using RDFa or Microdata. Creating GoodRelations snippets in RDFa or Microdata using the product information extracted from Web is a business value, especially when you consider most of the popular search engines recommend the use of these standards to provide rich site data for their index.

Keywords: product, GoodRelations, Protégé, RDFa, Microdata

1 Introduction

The Web contains a huge number of online sources which provides excellent resources for product information including specifications and descriptions of products. If we present this product information in a structured way, it will significantly improve the effectiveness of many applications [1]. This paper introduces IRIS wrapper to solve an increasingly important problem: extracting information from the product descriptions provided by online sources and structuring this information so that is sharable among business entities, software agents and search engines. The information extraction systems can be divided into three categories [2]: (a) *Procedural Wrapper:* The approach is based on writing customized wrappers for accessing required data from a given set of information sources. The extraction rules are coded into the program. Creating wrappers are easier and it can directly output the domain data model of application but each wrapper works only for an individual page. (b) *Declarative Wrapper:* These systems consist of a general execution engine and declarative extraction rules developed for specific data sources. The wrapper takes an input specification that declaratively states where the data of interest is located on the HTML document, and how the data should be wrapped into a new data model. (c) *Automatic Wrapper:* The automatic extraction approach uses machine learning techniques to learn extraction rules by examples. In [3] information extraction systems are classified

into two: solutions treating Web pages as a *tree*, and solutions treating Web pages as *data stream*. Systems are also divided with respect to the level of automation of wrapper creation into *manual*, *semi-automatic* and *automatic*. IRIS is a declarative and manual tree wrapper [1], which has a general rule engine that executes the rules specified in a template file using XML Path Language (XPath). Manual approaches are known to be tedious, time-consuming and require some level of expertise concerning the wrapper language [4]. However, manual and semi-automatic approaches are currently better suited for creating robust wrappers than the automatic approach. Writing an IRIS template is considerably easier than most of the existing manual wrappers. Besides, it can be predicted that to improve the reusability and the efficiency, the users of the IRIS engine will share templates on the Web.

There are works which directly focus on the problem of this paper. [5] uses a template-independent approach to extract product attribute name and value pair from Web. This approach makes hypothesis to identify the specification block but since some detail product pages may violate these hypothesis, the pairs in these pages cannot be extracted properly. The second work [6] needs two predefined ontologies to extract product attribute name and value pairs from a Web page. One of these ontologies is built according to the contents of the page but it is not an easy task to build that ontology from scratch for every change in the page content. The system presented in this paper differs from the above works in many ways.

First of all the system transforms the extracted information into an ontology to share and reuse common understanding of structure of information among users or software agents. To my knowledge [7], IRIS is the first Protégé plug-in that is used to extract product information from Web pages. Designed as a plug-in for the open source ontology editor Protégé, IRIS exploits the advantages of the ontology as a formal model for the domain knowledge and profits from the benefits of a large user community (currently *230,914* registered users).

Another feature is support for building an ontology that is compatible with GoodRelations Vocabulary [8], which is the most powerful vocabulary for publishing all of the details of your products and services in a way friendly to search engines, mobile applications, and browser extensions. The goal is to have extremely deep information on millions of products, providing a resource that can be plugged into any e-commerce system without limitation. If you have GoodRelations in your markup, Google, Bing, Yahoo, and Yandex will or plan to improve the rendering of your page directly in the search results. Besides, you provide information to the search engines so that they can rank up your page for queries to which your offer is a particularly relevant match. Finally, as an open source Java Application, IRIS can be further extended, fixed or modified according to the needs of the individual users.

The following section (with three subsections) includes the system's features and a scenario based quick-start guide. Section 3 concludes the paper with a brief talk about possible future work.

[1] Download link: https://github.com/tugbaozacar/iris

2 Scenario-based System Specification

IRIS system gathers semi-structured product information from an HTML page, applies extraction rules specified in the template file, and presents the extracted product data in an ontology that is compatible with GoodRelations Vocabulary. The HTML page is first parsed into a DOM tree using HtmlUnit, which is a Web Driver that supports walking the DOM model of the HTML document using XPath queries. In order to get product information from Web page, the template file includes a tree that specifies the paths of HTML tags around the product attribute names and product attribute values. Figure 1 shows the architecture of the system briefly. User builds a template for the pages containing the product information.

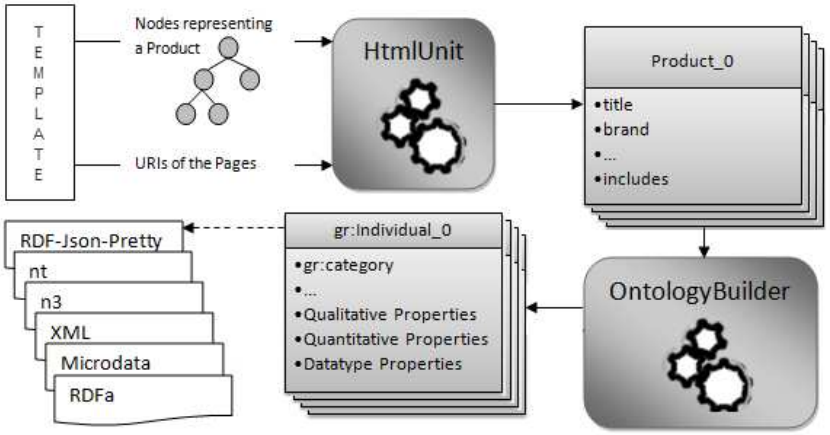


Fig. 1. Architecture of the system.

Then HtmlUnit library parses the Web pages. The system evaluates the nodes in the template and queries the HtmlUnit for the required product properties. At the end of this process, the system returns a list of product objects. To define a GoodRelations-compliant ontology the user maps the product properties to the properties of the "gr:Individual" class, saves the ontology and serializes the ontology into a series of structured data markup standards. The system makes serialization via RDF Translator API [9]). Each step is described in the following subsections.

2.1 Create a Template File

The information collected is mapped to the attributes of the *Product* object including title, description, brand, id, image, features, property names, property values and components. A template has two parts; the first part contains the tree that specifies the paths of HTML tags around the product attribute names and values. The second part specifies how

the HTML documents should be acquired. The product information is extracted using the tree. The tree is created manually and its nodes are converted to XPath expressions. HtmlUnit evaluates the specified XPath expressions and returns the matching elements.Figure 2 shows the example HTML code which contains the information about the first product in "amazon.com" pages that contain information about laptops. Figure 3 shows the tree which is built for extracting product information from the page in Figure 2.

```
<div id="result_0" class="fstRowGrid prod celwidget" name="B00D3F7H0K">
    <div class="linePlaceholder"></div>
    <div class="image imageContainer">
      <a href="http://www.amazon.com/Toshiba-Satellite-C55D-A5240 ... >
      <img  src="http://ecx.images-
      amazon.com/images/I/4108rtR%2BQWL. AA160_.jpg" class="productImage"
      alt="Product Details"/>                                  product.imgLink
      </a>
      <div class="smallVariationsBox"> </div>
    </div>
    <h3 class="newaps">
      <a href="http://www.amazon.com/Toshiba-Satellite-C55D-A524 ... >
       ...
      <span class="lrg bold">Toshiba Satellite C55D-A5240NR 15.6-Inch
      Laptop (Satin Black in Trax Horizon)
      </span>                                                  product.title
      </a>
    </h3>
    ...
</div>
```

Fig. 2. The HTML code which contains the first product on the page.

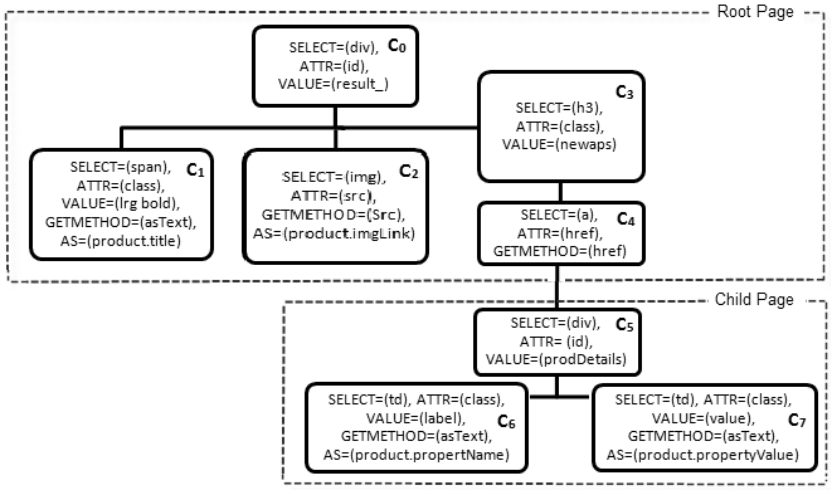


Fig. 3. The tree nodes in the template for "amazon.com" pages about laptops.

The leaf nodes of the tree (Figure 3) contains the HTML tag around a product attribute name or a product attribute value, and the internal nodes of the tree contains the HTML tags in which the HTML tag in the leaf node is nested. Therefore the hierarchy of the tree also represents the hierarchy of the HTML tags. c_1 contains the value of the title attribute, c_2 contains the image link of the product, and c_3 is one of the internal nodes that specify the path to its leaf nodes. c_3 specifies that all of its children contain HTML tags which are nested within the *h3* heading tag having class name "newaps". Its child node (c_4) specifies the HTML link element which goes to another Web page that contains detailed information about the product. The starting Web page is referred as root page and the pages navigated from root page are child pages. After jumping the page address specified by c_4, product properties and their values are chosen from this Web page which is shown in Figure 4.

```
<div id="prodDetails">
...
<tr><td class="label">Screen Size</td>
<td class="value">15.6 inches</td></tr>
<tr><td class="label">Screen Resolution</td>
<td class="value">1366 x 768</td></tr>
<tr><td class="label">Max Screen Resolution</td>
<td class="value">1366x768 pixels</td></tr>
...
</div>
```

Fig. 4. Product properties and their values.

The properties and their corresponding values are stored in an HTML table, which is nested in an HTML division identified by "prodDetails" *id*. Therefore c_5 specifies this HTML division and its child nodes c_6 and c_7 specifies the HTML cells containing product properties and their values. After determining the HTML elements which contain the product information, the user defines these elements in the template properly. Each node in the tree is a combination of the following fields:
SELECT-ATTR-VALUE These three fields are used to build the XPath query that specifies the HTML element in the page.
ORDER is used when there is more than one HTML element matching with the expression. The numeric value of the ORDER element specifies which element will be selected.
GETMETHOD is used to collect the proper values in the selected HTML element e. If you want to get the textual representation of the element (e), in other words what would be visible if this page was shown in a Web browser, you define the value of GETMETHOD field as "asText". Otherwise you get the value of an element (e) attribute by specifying the name of the attribute as the value of GETMETHOD field.
AS is only used with leaf nodes. The value collected from a leaf node using GETMETHOD field is mapped to the *Product* attribute specified in the AS field.

Appendix A gives the template (amazon.txt) which contains the code of the tree in Figure 3. The second part of a template file contains the information on how the HTML documents should be acquired. This part has the following fields:
NEXT_PAGE The information about laptops in "amazon.com" is spread across 400 pages. The link of the next page is stored in this field.
PAGE_RANGE specifies the number of the page or the range of pages which you want to collect information from. In my example, I want to collect the products in pages from 1 to 3.
BASE_URI represents the base URI of the site. In my example, the value of this field is http://www.amazon.com.
PAGE_URI is the URI of the first page which you want to collect information from. In my example, this is the URI of the page 1.
CLASS contains the name of the class that represents the products to be collected. In my example, "Laptop" class is used.

2.2 Create an Ontology that is Compatible with GoodRelations Vocabulary

First of all, user opens an empty ontology ("myOwl.owl") in the Protégé Ontology Editor and displays the IRIS tab which is listed on the TabWidgets panel. Then the user selects the template file using "Open template" button in Figure 5 (for this example: amazon.txt). Then the tool imports all laptops from the "amazon.com" pages specified in the PAGE_RANGE field. The imported individuals are listed in the "Individuals Window" (Figure 5). The "Properties Window" lists all properties of the individuals in "Individuals Window".
In this section, I follow up the descriptions and examples introduced in GoodRelations Primer [10]. First of all, the system defines the class in your template ("Laptop" class in example) as a subclass of "gr:Individual" class of the GoodRelations vocabulary. Then the properties of the "Laptop" class, which are collected from the Web page should be mapped to the properties of "gr:Individual", which can be classified as follows:
First category: "gr:category", "gr:color", "gr:condition", etc. (see [10] for full list). If the property p_x is semantically equivalent of a property from the first category p_y , then user simply maps p_x to p_y.
Second category: Properties that specify quantitative characteristics, for which an interval is at least theoretically an appropriate value should be defined as subproperties of "gr:quantitativeProductOrServiceProperty". A quantitative value is to be interpreted in combination with the respective unit of measurement and mostly quantitative values are intervals.
Third category: All properties for which value instances are specified are subproperties of "gr:qualitativeProductOrServiceProperty".
Fourth category: Only such properties that are no quantitative properties and that have no predefined value instances are defined as subproperties of "gr:datatypeProductOrServiceProperty".
To create a GoodRelations-compliant ontology, user selects the individuals and properties that will reside in the ontology. Then she clicks the "Use GoodRelations Vocabulary" button (Figure 5) and "Use GoodRelations Vocabulary" wizard appears. She selects the corresponding GoodRelations property type and respective unit of measurement.

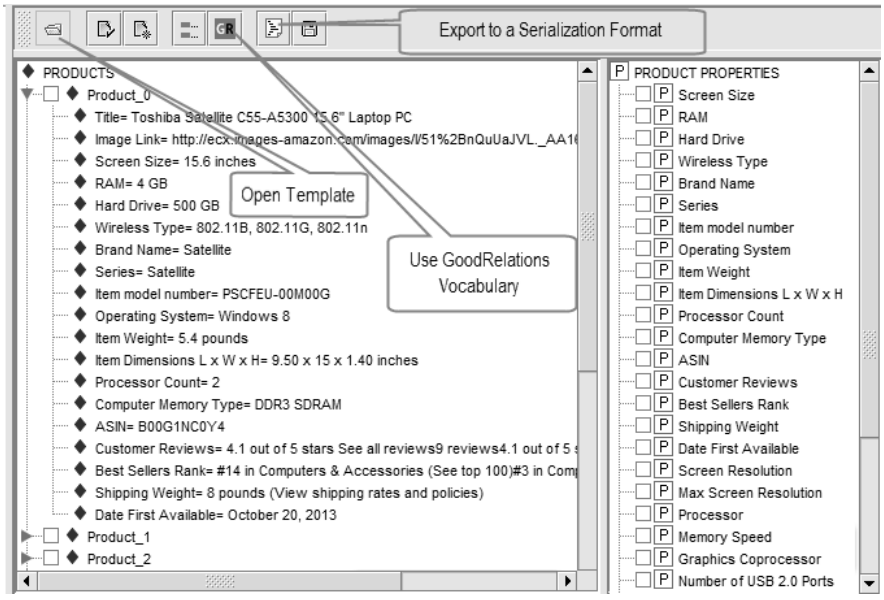


Fig. 5. The tool imports all laptops from the specified "amazon.com" pages.

2.3 Save and Serialize the Ontology

User saves the ontology in an owl file and clicks the "Export to a serialization format" button (Figure 5) to view the ontology in one of the structured data markup standards.

3 Conclusion and Future Work

This work introduces a Protégé plug-in called IRIS that collects product information from Web and transforms this information into GoodRelations snippets in RDFa or Microformats. The system attempts to solve an increasingly important problem: extracting useful information from the product descriptions provided by the sellers and structuring this information into a common and sharable format among business entities, software agents and search engines. I plan to improve the IRIS plug-in with an extension that gets user queries and sends them to Semantics3 API [11], which is a direct replacement for Google's Shopping API and gives developers comprehensive access to data across millions of products and prices. Another potential future work is generating an environment for semi-automatic template construction. An environment that automatically constructs the tree nodes from the selected HTML parts will significantly reduce the time to build a template file. And yet another future work is diversify the supported input formats (pdf, excel, csv etc.).

Appendix A

```
SELECT=(div), ATTR=(id), VALUE= (result_) [
  SELECT=(span), ATTR=(class), VALUE=(lrg bold),
  GETMETHOD=(asText, AS=(product.title);
  SELECT=(img), ATTR=(src), GETMETHOD=(Src),
  AS=(product.imgLink);
  SELECT=(h3), ATTR=(class), VALUE= (newaps) [
    SELECT=(a), ATTR=(href), GETMETHOD=(href) [
      SELECT=(div), ATTR=(id), VALUE=(prodDetails)[
        SELECT=(td), ATTR=(class), VALUE=(label),
        GETMETHOD=(asText, AS=(product.propertyName);
        SELECT=(td), ATTR=(class), VALUE=(value),
        GETMETHOD=(asText, AS=(product.propertyValue)]]]]
NEXT_PAGE:{SELECT=(a), ATTR=(id), VALUE=(pagnNextLink),
           GETMETHOD=(href)}
PAGE_RANGE:{1-3}
BASE_URI:{http://www.amazon.com}
PAGE_URI:{http://www.amazon.com/s/ref=sr\_nr\_n\_1?rh=
          n\%3A565108\%2Ck\%3Alaptop\&keywords=laptop\&
          ie=UTF8\&qid=1374832151\&rnid=2941120011}
CLASS:{Laptop}
```

References

1. Tang, W., Hong, Y., Feng, Y.H., Yao, J.M., Zhu, Q.M.: Simultaneous product attribute name and value extraction with adaptively learnt templates. In: Proceedings of CSSS '12. (2012) 2021–2025
2. Han, J.: Design of Web Semantic Integration System. PhD thesis, Tennessee State University. (2008)
3. Firat, A.: Information Integration Using Contextual Knowledge and Ontology Merging. PhD thesis, MIT, Sloan School of Management (2003)
4. Muslea, I., Minton, S., Knoblock, C.: A hierarchical approach to wrapper induction, ACM Press (1999) 190–197
5. Wu, B., Cheng, X., Wang, Y., Guo, Y., Song, L.: Simultaneous product attribute name and value extraction from web pages. In: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference, IEEE Computer Society (2009) 295–298
6. Holzinger, W., Kruepl, B., Herzog, M.: Using ontologies for extracting product features from web pages. In: Proceedings of the ISWC'06, Springer-Verlag 2006 (2006) 286–299
7. : Protege plug-in library Last accessed: 2013-09-24.
8. Hepp, M.: Goodrelations: An ontology for describing products and services offers on the web. EKAW '08 (2008) 329–346
9. Stolz, A., Castro, B., Hepp, M.: Rdf translator: A restful multiformat data converter for the semantic web. Technical report, E-Business and Web Science Research Group (2013)
10. Hepp, M.: Goodrelations: An ontology for describing web offers —primer and user's guide. Technical report, E-Business + Web Science Research Group 2008.
11. Semantics3 Inc.: Semantics3 - apis for products and prices (2013)

Ontology Design Patterns: Adoption Challenges and Solutions

Karl Hammar

Jönköping University
P.O. Box 1026
551 11 Jönköping, Sweden
karl.hammar@jth.hj.se

Abstract. Ontology Design Patterns (ODPs) are intended to guide non-experts in performing ontology engineering tasks successfully. While being the topic of significant research efforts, the uptake of these ideas outside the academic community is limited. This paper summarises some issues preventing broader adoption of Ontology Design Patterns among practitioners, suggests research directions that may help overcome these issues, and presents early results of work in these directions.

Keywords: Ontology Design Pattern, eXtreme Design, Tools

1 Introduction

Ontology Design Patterns (ODPs) were introduced by Gangemi [8] and Blomqvist & Sandkuhl [4] in 2005 (extending upon ideas by the W3C Semantic Web Best Practices and Deployment Working Group[1]), as a means of facilitating practical ontology development. These patterns are intended to help guide ontology engineering work, by packaging best practice into small reusable blocks of ontology functionality, to be adapted and specialised by users in individual ontology development use cases.

This idea has gained some traction within the academic community, as evidenced by the Workshop on Ontology Patterns series of workshops held on conjunction with the International Semantic Web Conference. However, the adoption of ODPs among practitioners is still quite limited. If such patterns are to be accepted as useful artefacts also in practice, it is essential that they [10]:

- model concepts and phenomena that are relevant to practitioners' needs
- are constructed and documented in a manner which makes them accessible and easy to use by said practitioners in real-world use cases
- are accompanied by appropriate methods and tools that support their use by the intended practitioners

[1] http://www.w3.org/2001/sw/BestPractices/

While the first requirement above can be said to be fulfilled by the ODPs published online (the majority of which result from projects and research involving both researchers and practitioners), the latter two requirements have largely been overlooked by the academic community. Many patterns are poorly documented, and at the time of writing, none have been sufficiently vetted to graduate from *submitted* to *published* status in the prime pattern repository online[2]. Toolset support is limited to some of the tasks required when employing patterns, while other tasks are entirely unsupported. Furthermore, the most mature pattern usage support tools are implemented as a plugin for an ontology engineering environment which is no longer actively maintained[3].

In the following paper, these ODP adoption challenges are discussed in more detail, and the author's ongoing work on addressing them is reported. The paper focuses exclusively on Content ODPs as defined in the NeOn Project[4], as this is most common type of Ontology Design Patterns with some 100+ patterns published. The paper is structured as follows: Section 2 introduces relevant related published research on ODPs, Section 3 focuses on the tasks that need be performed when finding, adapting, and applying patterns, Section 4 details the challenges preventing the adoption of ODPs by practitioner ontologists, Section 5 proposes solutions to these challenges, Section 6 presents the initial results of applying some of those solutions, and Section 7 concludes and summarises the paper.

2 Related Work

Ontology Design Patterns were introduced as potential solutions to these types of issues at around the same time independently by Gangemi [8] and Blomqvist and Sandkuhl [4]. The former define such patterns by way of a number of characteristics that they display, including examples such as *"[an ODP] is a template to represent, and possibly solve, a modelling problem"* [8, p. 267] and *"[an ODP] can/should be used to describe a 'best practice' of modelling"* [8, p. 268]. The latter describes ODPs as generic descriptions of recurring constructs in ontologies, which can be used to construct components or modules of an ontology. Both approaches emphasise that patterns, in order to be easily reusable, need to include not only textual descriptions of the modelling issue or best practice, but also some formal ontology language encoding of the proposed solution. The documentation portion of the pattern should be structured and contain those fields or slots that are required for finding and using the pattern.

Since their introduction, ODPs have been the subject of some research and work, see for instance the deliverables of the EU FP6 NeOn Project[5] [15, 5] and the work presented at instances of the Workshop on Ontology Patterns[6]

[2] `http://ontologydesignpatterns.org/`
[3] XD Tools for NeOn Toolkit, `http://neon-toolkit.org/wiki/XDTools`
[4] `http://ontologydesignpatterns.org/wiki/Category:ContentOP`
[5] `http://www.neon-project.org/`
[6] `http://ontologydesignpatterns.org/wiki/WOP:Main`

at the International Semantic Web Conference. There are to the author's best knowledge no studies indicating ontology engineering performance improvements in terms of time required when using patterns, but results so far indicate that their usage can help lower the number of modelling errors and inconsistencies in ontologies, and that they are perceived as useful and helpful by non-expert users [3, 6].

The use and understanding of ODPs have been heavily influenced by the work taking place in the NeOn Project[7], the results of which include a pattern typology [15], and the eXtreme Design collaborative ontology development methods, based on pattern use [5]. eXtreme Design (XD) is defined as *"a family of methods and associated tools, based on the application, exploitation, and definition of Ontology Design Patterns (ODPs) for solving ontology development issues"* [14, p. 83]. The method is influenced by the eXtreme Programming (XP)[2] agile software development method, and like it, emphasises incremental development, test driven development, refactoring, and a divide-and-conquer approach to problem-solving [13]. Additionally, the NeOn project funded the development of the XD Tools, a set of plugin tools for the NeOn Toolkit IDE intended to support the XD method of pattern use.

Ontology Design Patterns have also been studied within the CO-ODE project[1, 7], the results of which include a repository of patterns[8] and an Ontology Pre-Processing Language (OPPL)[9].

3 Using Ontology Design Patterns

The eXtreme Design method provides recommendations on how one should structure an Ontology Engineering project of non-trivial size, from tasks and processes of larger granularity (project initialisation, requirements elicitation, etc) all the way down to the level of which specific tasks need be performed when employing a pattern to solve a modelling problem. Those specific pattern usage tasks (which are also applicable in other pattern-using development methods) are:

1. Finding patterns relevant to the particular modelling issue
2. Adapting those general patterns to the modelling use case
3. Integrating the resulting specialisation with the existing ontology (i.e., the one being built)

3.1 Finding ODPs

In XD, the task of finding an appropriate design pattern for a particular problem is viewed as a matching problem where a local use case (the problem for which the ontology engineer needs guidance) is matched to a general use case

[7] `http://www.neon-project.org/`
[8] `http://odps.sourceforge.net/odp/html/index.html`
[9] `http://oppl2.sourceforge.net/`

(the intended functionality of the pattern) encoded in the appropriate pattern's documentation. In order to perform such matching, the general use case needs be expressed in a way that enables matching to take place. In practice, pattern intent is encoded using Competency Questions [9], and matching is performed by hand, by the ontology engineer him/herself. XD Tools supports rudimentary keyword-based search across the ontologydesignpatterns.org portal, which can provide the ontology engineer with an initial list of candidate patterns for a given query.

3.2 Specialising ODPs

Having located a pattern appropriate for reuse in a specific scenario, the ontology engineer needs to adapt and specialise said pattern for the scenario in question. The specific steps vary from case to case, but a general approach that works in the majority of cases is as follows:

1. Specialise leaf classes of the subclass tree
2. Specialise leaf properties of the subproperty tree
3. Define domains and ranges of specialised properties to correspond with the specialised classes

The XD Tools provide a wizard interface that supports each these steps. They also provide a certain degree of validation of the generated specialisations, by presenting the user with a list of generated axioms (expressed in natural language) for the user to reject or accept.

3.3 Integrating ODP Instantiations

Once a pattern has been adapted for use in a particular scenario, the resulting solution module needs to be integrated with the ontology under development. This integration involves aligning classes and properties in the pattern module with existing classes and properties in the ontology, using subsumption or equivalency mappings. This integration process may also include refactoring of the existing ontology, in the case that requirements dictate that the resulting ontology be highly harmonised. There is at the time of writing no known tool support for ODP instantiation integration, and this process is therefore performed entirely by hand.

4 ODP Adoption Challenges

As indicated above, there is a thriving research community studying patterns and developing new candidate ODPs. Unfortunately the adoption of Ontology Design Patterns in the broader Semantic Web community, and in particular among practitioners, is limited. The author has, based on experiences from several studies involving users on different levels (from graduate students to domain

experts from industry) [12, 10, 11], identified a number of issues that give rise to confusion and irritation among users attempting to employ ODPs, and which are likely to slow uptake of these technologies. Those issues are detailed in the subsequent sections.

4.1 Issues on Finding ODPs

As explained, there are two methods for finding appropriate design patterns for a particular modelling challenge - users can do matching by hand (by consulting a pattern repository and reading pattern documentations one by one), or users can employ the pattern search engine included in XD Tools to suggest candidate patterns. In the former case, as soon as the list of available patterns grows to a non-trivial number (such as in the ontologydesignpatterns.org community portal), users find the task challenging to perform correctly, particularly if patterns are not structured in a way that is consistent with their expectations [10].

In the latter case, signal-to-noise ratio of pattern search engine results is often discouragingly low. In initial experiments (detailed in Section 6) the author found that with a result list displaying 25 candidate patterns, the correct pattern was included in less than a third of the cases. In order to guarantee that the correct pattern was included, the search engine had to return more than half of the patterns in the portal, essentially negating the point of using a search engine. Also, the existing pattern search engine included in XD Tools does not allow for filtering the results based on user criteria, which makes it easy for a user to mistakenly import and apply a pattern which is inconsistent with ontology requirements, e.g., on reasoning performance or other constraints.

4.2 Issues on Composing ODPs

The process of integrating a specialised pattern solution module into the target ontology is not supported by any published tools, and consequently relies entirely on the user's ontology engineering skill. Users performing such tasks are often confused by the many choices open to them, and the potential consequences of these choices, not limited to:

- Which mapping axioms should be used between the existing classes and properties and those of the solution module, e.g., equivalency or subsumption?
- Where those pattern instantiation module mapping axioms should be placed: in the target ontology, in the instantiated pattern module, or in a separate mapping module?
- The interoperability effects of customising patterns: for instance, what are the risks in case pattern classes are declared to be subsumed by existing top level classes in the target ontology?
- How selections from the above composition choices affect existing ontology characteristics such as reasoning performance, etc.

4.3 Issues on Pattern and Tooling Quality

Users often express dissatisfaction with the varying degree of documentation quality [10]. While some patterns are documented in an exemplary fashion, many lack descriptions of intents and purpose, consequences of use, or example use cases. Experienced ontology engineers can see through this by studying the accompanying OWL module in order to learn the benefits and drawbacks of a certain pattern, but it is uncommon for non-expert users to do this successfully.

It is not uncommon for patterns to include and build upon other patterns, and these dependencies are not necessarily intuitive or well-explained. On several occasions the author has been questioned by practitioner users as to why, in the `ontologydesignpatterns.org` repository, the pattern concerning time indexed events makes use of the *Event* class that is defined in the (non time-indexed) *Participation* pattern. The consequence of this dependency structure is of course that any user who models time indexed events using patterns automatically also includes non time-indexed participation representations in their resulting model, which very easily gives rise to modelling mistakes.

In more practical terms, the XD Tools were designed to run as a plugin for the NeOn Toolkit ontology IDE. This IDE unfortunately never gained greater adoption. Additionally, XD Tools and its dependencies require a specific older version of NeOn Toolkit. This means that ontology engineers who want to use newer tools and standards are unable to use XD Tools, but rather have to do their pattern-based ontology engineering without adequate tool support.

5 Improvement Ideas

The author's ongoing research aims to improve upon ODP usage methods and tools, in the process solving some of the issues presented above. To this end, a number of solution suggestions have been developed, and are currently in the process of being tested (some with positive results, see Section 6). The following sections present these suggestions and the consequences they would have on both patterns and pattern repositories. Implementation of these suggested improvements within an updated version of the XD Tools targeting the Protégé editor is planned to take place in the coming months.

5.1 Improving ODP Findability

In order to improve recall when searching for suitable ODPs, the author suggests making use of two pieces of knowledge regarding patterns that the current XD Tools pattern search engine does not consider: firstly, that the core intent of the patterns in the index is codified as competency questions, which are structurally similar to such queries that an end-user may pose, and secondly, that patterns are general or abstract solutions to a common problem, and consequently, the specific query that a user inputs needs to be transformed into a more general form in order to match the indexed patterns level of abstraction.

The first piece of knowledge can be exploited by using string distance metrics to determine how similar an input query is to the competency questions associated with a pattern solution. Another approach under study is to employ ontology learning methods to generate graphs from both indexed pattern competency questions and input queries, and then measuring the degree of overlap between concepts referenced in these two graphs.

The second piece of knowledge can be exploited by reusing existing language resources that represent hyponymic relations, such as WordNet. By enriching the indexed patterns with synonyms of disambiguated classes and properties in the pattern, and by enriching the user query using hypernym terms of the query, the degree of overlap between a user query (worded to concern a specific modelling issue) against a pattern competency question (worded to concern a more general phenomenon) can be computed.

5.2 Improving ODP Integration

The challenge of integrating an instantiated pattern module into a target ontology is at its core an ontology alignment challenge. Consequently existing ontology alignment and ontology matching methods are likely to be useful in this context. The behaviour of such systems against very small ontologies such as instantiated pattern modules, is however not well known. The advantage that patterns have over general ontologies in this context is the knowledge that patterns are designed with the very purpose of being adapted and integrated into other ontologies, which is not true in the general ontology alignment use case. Therefore, the pattern creator could a priori consider different ways in which that pattern would best be integrated with an ontology, and construct the pattern in such a way as to make this behaviour known to an alignment system.

The author suggests reusing known good practice from the ontology alignment domain, and combining this with such pattern-specific alignment hints embedded in the individual pattern OWL files. For instance, a pattern class could be tagged with an annotation indicating to a compatible alignment system that this class represents a very high level or foundational concept, and that consequently, it should not be aligned as a subclass; or a pattern class or property could be tagged with annotations indicating labels of suitable sub- or superclasses in the integration step.

Additionally, improved user interfaces would aid non-expert users in applying patterns. Such user interfaces should detail in a graphical or otherwise intuitive manner the consequences of selecting a particular integration strategy, in the case that multiple such strategies are available for consideration.

6 Results

The author has developed a method of indexing and searching over a set of Ontology Design Patterns based on the ideas presented in Section 5. The method combines the existing Lucene-backed Semantic Vectors Search method with a

comparison of competency questions based on their relative Levenshtein edit distances, and a comparison of the number of query hypernyms that can be found among the pattern concept synonyms. Each method generates a confidence value between 0 and 1, and these confidence values are added together with equal weight to generate the final confidence value which is used for candidate pattern ordering. While the approach requires further work, early results are promising, as shown in Table 1.

The dataset used in testing was created by reusing the question sets provided by the *Question Answering over Linked Data* (QALD) evaluation campaign. Each question was matched to one or more ODPs suitable for building an ontology supporting the question. This matching was performed by two senior ontology experts independently, and their respective answer sets merged. The two experts reported very similar pattern selections in the cases where only a single pattern candidate existed in the pattern repository compliant with a competency question (e.g., the *Place*[10] or *Information Realization*[11] patterns), but for such competency questions where multiple candidate patterns existed representing different modelling practices (e.g., the *Agent Role*[12] or *Participant Role*[13] patterns), their selections among these candidate patterns diverged. Consequently, the joint testing dataset was constructed via the union of the two experts' pattern selections (representing the possibility of multiple correct modelling choices), rather than their intersection. Recall was defined as the ratio of such expert-provided ODP candidates that the automated system retrieves for a given input question.

Table 1. Recall Improvement for ODP Search

	XD-SVS	Composite3
R10	6 %	22 %
R15	8 %	31 %
R20	9 %	37 %
R25	14 %	41 %

As shown in the table, the average recall within the first 10, 15, 20 or 25 results is 3-4 times better using the author's composite method (Composite3) than using the existing XD Tools Semantic Vectors Search (XD-SVS). It should be noted that while Composite3 also increases the precision of the results compared to XD-SVS by a similar degree, that resulting precision is still rather poor, at 5-6 %. The potential pattern user will consequently see a lot of spurious results

[10] `http://ontologydesignpatterns.org/wiki/Submissions:Place`
[11] `http://ontologydesignpatterns.org/wiki/Submissions:Information\_realization`
[12] `http://ontologydesignpatterns.org/wiki/Submissions:AgentRole`
[13] `http://ontologydesignpatterns.org/wiki/Submissions:ParticipantRole`

using either of the approaches. This is understood to be a potential usability problem, and an area for further work.

A factor believed to be limiting the success of this method is the fact that resolving ODP concepts and properties to corresponding concepts and properties in natural language resources (in this case WordNet) is an error-prone process. This is largely due to the ambiguity of language and the fact that concepts in ODPs are generally described using only a single label per supported language. If pattern concepts were more thoroughly documented, using for instance more synonymous labels, class sense disambiguation would likely work better, and ODP search consequently work better also. Additionally, WordNet does contain parts of questionable quality (both in terms of coverage and structure), the improvement of which may lead to increased quality of results for dependent methods such as the one presented here.

7 Conclusions

This paper has introduced and discussed some concrete challenges regarding the use of Ontology Design Patterns, with an emphasis on tooling-related challenges that prevent non-expert users from performing Ontology Engineering using such patterns. Those challenges primarily concern; a) the task of finding patterns, b) decisions to make when integrating pattern based modules with an existing ontology, and, c) pattern and tooling quality. The author's work aims to overcome these challenges by developing improved methods and accompanying tools for today's Ontology Engineering IDE:s (i.e., Protégé), better supporting each step of ODP application and use.

The author has developed an ODP search method exploiting both the similarity between pattern competency questions and user queries, and the relative abstraction level of general pattern solutions versus concrete user queries, a method shown to increase recall when searching for candidate ODPs significantly. Future work includes improving recall and precision further, and developing methods and tooling to support the ODP integration task.

References

1. Aranguren, M.E., Antezana, E., Kuiper, M., Stevens, R.: Ontology Design Patterns for Bio-ontologies: A Case Study on the Cell Cycle Ontology. BMC bioinformatics 9(Suppl 5), S1 (2008)
2. Beck, K., Andres, C.: Extreme Programming Explained: Embrace Change. Addison-Wesley Professional (2004)
3. Blomqvist, E., Gangemi, A., Presutti, V.: Experiments on Pattern-based Ontology Design. In: Proceedings of the Fifth International Conference on Knowledge Capture. pp. 41–48. ACM (2009)
4. Blomqvist, E., Sandkuhl, K.: Patterns in Ontology Engineering: Classification of Ontology Patterns. In: Proceedings of the 7th International Conference on Enterprise Information Systems. pp. 413–416 (2005)

5. Daga, E., Blomqvist, E., Gangemi, A., Montiel, E., Nikitina, N., Presutti, V., Villazon-Terrazas, B.: D2.5.2: Pattern Based Ontology Design: Methodology and Software Support. Tech. rep., NeOn Project (2007)
6. Dzbor, M., Suárez-Figueroa, M.C., Blomqvist, E., Lewen, H., Espinoza, M., Gómez-Pérez, A., Palma, R.: D5.6.2 Experimentation and Evaluation of the NeOn Methodology. Tech. rep., NeOn Project (2007)
7. Egaña, M., Rector, A., Stevens, R., Antezana, E.: Applying Ontology Design Patterns in Bio-Ontologies. In: Knowledge Engineering: Practice and Patterns, pp. 7–16. Springer (2008)
8. Gangemi, A.: Ontology Design Patterns for Semantic Web Content. In: The Semantic Web–ISWC 2005, pp. 262–276. Springer (2005)
9. Grüninger, M., Fox, M.S.: The role of competency questions in enterprise engineering. In: Benchmarking—Theory and Practice, pp. 22–31. Springer (1995)
10. Hammar, K.: Ontology Design Patterns in Use: Lessons Learnt from an Ontology Engineering Case. In: Proceedings of the 3rd Workshop on Ontology Patterns (2012)
11. Hammar, K.: Towards an Ontology Design Pattern Quality Model (2013)
12. Hammar, K., Lin, F., Tarasov, V.: Information Reuse and Interoperability with Ontology Patterns and Linked Data. In: Business Information Systems Workshops. pp. 168–179. Springer (2010)
13. Presutti, V., Blomqvist, E., Daga, E., Gangemi, A.: Pattern-Based Ontology Design. In: Ontology Engineering in a Networked World, pp. 35–64. Springer (2012)
14. Presutti, V., Daga, E., Gangemi, A., Blomqvist, E.: eXtreme Design with Content Ontology Design Patterns. In: Proceedings of the Workshop on Ontology Patterns (WOP 2009), collocated with ISWC 2009. p. 83 (2009)
15. Presutti, V., Gangemi, A., David, S., Aguado de Cea, G., Suárez-Figueroa, M.C., Montiel-Ponsoda, E., Poveda, M.: D2.5.1: A Library of Ontology Design Patterns: Reusable Solutions for Collaborative Design of Networked Ontologies. Tech. rep., NeOn Project (2007)

Mapping Representation based on Meta-data and SPIN for Localization Workflows

Alan Meehan, Rob Brennan, Dave Lewis, Declan O'Sullivan

CNGL Centre for Global Intelligent Content, Knowledge and Data Engineering Group, School of Computer Science and Statistics, Trinity College Dublin, Ireland
{meehanal,rob.brennan,dave.lewis,declan.osullivan}@scss.tcd.ie

Abstract. The localization industry currently deploys language translation workflows based on heterogeneous tool-chains. Standardized tool interchange formats such as XLIFF (XML Localization Interchange File Format) have had some impact on enabling more agile translation workflows. However the rise of new tools based on machine translation technology and the growing demand for enterprise linked data applications has created new interoperability challenges as workflows need to encompass a broader range of tools. In this paper we present an approach of representing mappings between RDF-based representations of multilingual content and meta-data. To represent the mappings, we use a combination of SPARQL Inferencing Notation (SPIN) and meta-data. Our approach allows the mapping representation to be published as Linked Data. In contrast to other frameworks such as R2R, the mappings are executed via a standard SPARQL processor. The objective is to provide a more agile approach to translation workflows and greater interoperability between software tools by leveraging the ongoing innovation in the Multilingual Web field. Our use case is a Language Technology retraining workflow where publishing mappings leads to new opportunities for interoperability and end-to-end tool-chain analytics. We present the results from an initial experiment which compared our approach of executing and representing mappings to that of a similar approach - the R2R Framework.

Keywords: Multilingual Web, Semantic Mapping, Interoperability

1 Introduction

The localization industry is historically built on fragile cross-enterprise tool-chains with strong interoperability requirements. Ongoing research and innovation by the Multilingual Semantic Web and Linked Data communities has led to promising new technologies that can simultaneously span the language and interoperability barriers. However switching to an enterprise Linked Data model is not a straight forward task. Data needs to be transformed from its original format into a RDF-based representation and multiple domain or tool-specific vocabularies are often employed within the RDF. This increases the importance of mapping technology [1] to enable flexible end-to-end tool-chains. Where such tool-chains are in place, there is considerable com-

mercial advantage to enabling end-to-end analytics that can monitor content flows through the tools and the impact of mapping steps.

This paper focuses on the problem of making such mapping steps visible within a localization tool-chain, exposing the mappings in a way that facilities discovery, lifecycle management and the recording of mapping meta-data such as the mapping provenance. These mappings must be executable in the sense that it is desirable to have a framework that takes the mapping representation and can apply it as needed to instance data. By avoiding proprietary technologies in the execution step it is hoped that a wider range of tool vendors can be used to lower costs and simplify integration across multiple enterprises in a localization value chain.

This leads to the following two research questions that are investigated in this paper. How can mappings be expressed as Linked Data to facilitate discovery and the recording of mapping meta-data? To what extent can standard SPARQL endpoints act as an execution engine for these mappings?

We represent the executable RDF-to-RDF mappings as SPARQL[1] construct queries that can be executed on any standard SPARQL endpoint. To support mapping publication, discovery and meta-data annotation we represent the mappings as SPARQL Inference Notation (SPIN) [9] Linked Data. We aim to exploit this capability by also publishing associated mapping meta-data that will lead to new techniques for mapping lifecycle management and SPARQL-based mapping quality analytics.

Although a work in progress, the contribution of this paper is an evaluation of the relative expressivity of representing executable mappings as SPARQL construct queries compared to the mapping language of the R2R Framework [8] based on a set of test mappings previously published by the R2R team. In addition the viability of using SPIN to publish the mappings as Linked Data is evaluated by transforming the test mapping set into SPIN-based RDF representations.

The remainder of the paper is as follows: Section 2 presents a use case to illustrate where our approach of mapping representation would be useful; Section 3 presents the requirements of the mapping representation; Section 4 covers related work in the area of semantic mapping and the publication of mappings; Section 5 presents the evaluation of our approach of representing and executing mappings against the R2R Framework; we finish with conclusions and future work in Section 6.

2 Language Technology Retraining Workflow Use Case

This section describes a use case centered on the localization industry's process of providing translated content. This was chosen as an exemplar of complex real world workflows that the authors were familiar with. We focus on a Language Technology (LT) retraining workflow[2], with the goal to provide a means for translated content to be retrieved and used to retrain multiple machine translation tools.

[1] http://www.w3.org/TR/sparql11-query/

[2] Currently an ongoing development at CNGL Centre for Global Intelligent Content: http://www.cngl.ie

Figure 1 illustrates the process whereby a piece of HTML source content undergoes a series of processing steps, acted on by specific tools for translation, quality assessment and post editing. An XML Localization Interchange File Format (XLIFF)[3] file is used to record the processing that the content has undergone at each step. At the end of the content flow, a custom tool, using the Extensible Stylesheet Language Transformation (XSLT)[4] language, is used to map the data from the XLIFF file into RDF using the Global Intelligent Content semantic model (GLOBIC)[5] vocabulary and stored in a triple store. This RDF data represents details such as the *source* and *target* of text content that underwent a Machine Translation (MT) process, which tool carried out the MT process, *post edits* and *quality estimates* associated with translated content. By building up data in the triple store, it becomes a rich source of MT training data. A high quality training data-set is important for MT applications in order to gain benefits during training phases.

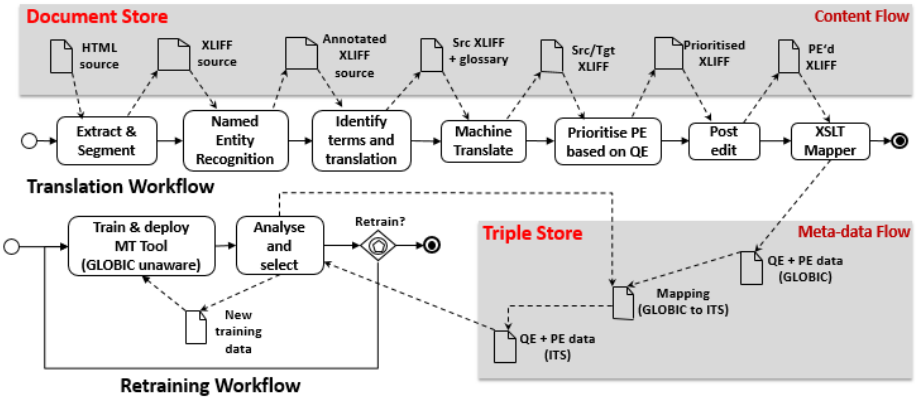


Figure 1. Language Technology Retraining Workflow

The retraining aspect of the workflow involves retrieving suitable content to be re-fed into the MT statistical machine learning tool. This is achieved by querying the triple store for translated content with a quality estimate over a certain threshold value, which is easily achieved using SPARQL queries.

Heterogeneous tools looking to utilize this training data naturally need to have the data in the triple store mapped to a schema they recognize. In Figure 1, the *MT tool* is GLOBIC unaware, it is designed to use the content represented by the Internationalization Tag Set[6] (ITS) vocabulary. Thus the *Quality Estimate (QE)* and *Post Edited (PE) data* that is represented in GLOBIC must be mapped to an ITS representation for

[3] http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html

[4] http://www.w3.org/TR/xslt

[5] The GLOBIC semantic model is an ongoing development at CNGL Centre for Global Intelligent Content. Its purpose in this use case is to enable greater interoperability and analytics within the workflow: http://www.scss.tcd.ie/~meehanal/gic.ttl

[6] http://www.w3.org/TR/its20/

the *MT tool* to use it. As will be seen in Section 3, our approach to representing such mappings allows them to be published alongside the other data in the triple store. This allows the mappings to be discovered by users/tools through SPARQL queries and executed by the SPARQL processor itself when transformed back to SPARQL syntax from SPIN. Transforming from SPARQL syntax to SPIN or vice-versa can be can be done using the SPIN RDF Converter[7], which can be used as a free web service.

3 Mapping Representation Requirements and Design

This section describes the requirements for the mapping representation and an example of how a mapping is represented under our approach.
The requirements of the mapping representation are as follows:

1. A mapping entity must be expressed as RDF, with a unique URI, allowing the mapping to be publishable on the web and discoverable via SPARQL queries.
2. The executable mapping statement must be a SPARQL query that is executable by a SPARQL processor.
3. The executable mapping statement must be expressed as RDF and must have a unique URI, allowing the statement to be queried by SPARQL and linked to a mapping entity.
4. A mapping entity is to be modeled with associated meta-data expressed as RDF, providing additional data on the mapping which can be queried via SPARQL.

To fulfil *requirement 1*, a mapping entity should be given a meaningful, unique name and modeled as an instance of the *Mapping* class from the GLOBIC vocabulary. For *requirement 2*, the executable mapping statement should be devised as a SPARQL construct query. For *requirement 3*, the SPARQL construct query should be converted to SPIN in order to be expressed as RDF and given a unique, meaningful name. The *hasRepresentation* property from the GLOBIC vocabulary should be used to link the SPIN representation of the SPARQL construct query to the mapping entity. For *requirement 4*, the associated meta-data for a mapping entity should be modeled using the following properties. The *wasCreatedBy*, the *mapDescription* and the *version* properties from the GLOBIC vocabulary and the *generatedAtTime* and *wasRevisionOf* properties from the W3C PROV[8] vocabulary are used.

Below is an example of a mapping representation that concerns the mapping of a MT quality score from the GLOBIC vocabulary to the ITS vocabulary. The SPIN representation of the SPARQL construct query is as follows:

```
01: @PREFIX gic: <http://www.scss.tcd.ie/~meehanal/gic#>.
02: @PREFIX itsrdf: <http://www.w3.org/2005/11/its/rdf#>.
03: @PREFIX sp: <http://spinrdf.org/sp#>.
04: @PREFIX ex: <http://www.example.org/example#>.
```

[7] http://spinservices.org/spinrdfconverter.html

[8] The PROV ontology contains classes and properties that can be used to model provenance data about an entity, agent or activity: http://www.w3.org/TR/prov-o/

```
05: ex:globic_to_its_mtScore_sp_2 a sp:Construct;
06:  sp:templates ([ sp:object _:b1;
07:                  sp:predicate itsrdf:mtConfidence;
08:                  sp:subject _:b2 ]);
09:  sp:where ([ sp:object _:b1;
10:              sp:predicate gic:qualityAssessment;
11:              sp:subject _:b2 ]).
12: _:b2 sp:varName "s"^^xsd:string.
13: _:b1 sp:varName "val"^^xsd:string.
```

The mapping entity plus associated meta-data is as follows:

```
14: ex:globic_to_its_mtScore_map_2 a gic:Mapping;
15:  gic:hasRepresentation ex:globic_to_its_mtScore_sp_2;
16:  gic:wasCreatedBy ex:person_1;
17:  prov:generatedAtTime "2014-01-01"^^xsd:date;
18:  gic:mapDescription "Used to map X to Y etc…";
19:  gic:version "1.1"^^xsd:float;
20:  prov:wasRevisionOf ex:globic_to_its_mtScore_map_1.
```

Examining the example above, *line 14* contains the name of the mapping entity. *Line 15* links the mapping to the SPIN representation of the SPARQL construct query on *line 05*. *Line 16* indicates what person/application is responsible for creating the mapping. *Line 17* indicates when the mapping was created. *Line 18* provides a human readable description of what the mapping does. *Line 19* indicates the current version of the mapping. *Line 20* provides a link to the previous version of the mapping.

4 Related Work

There is a rich body of research in semantic mapping undertaken by the semantic web community [1]. A wide variety of approaches have been adopted to tackle the mapping challenge, from rule-based representations [2], to axiomatic representations [3], to SPARQL query representations [4-5].

Keeney et al. [6] evaluated these three mapping approaches and found that the SPARQL query approach in general excels in terms of execution time and efficient use of computational resources. Although there are some particular circumstances where there are downsides to this approach. Keeney et al. conclude that for tasks where applications wish to map and use relatively small, specific data, the SPARQL approach would be ideal.

Little research has been undertaken into publishing mappings in order for them to be discovered and re-used. Thomas et al. [7] propose that the lack of ontology and mapping meta-data impedes the task of discovering relevant mappings between ontologies. They propose a thirty-three element mapping meta-data ontology, *OM²R*, based on the mapping lifecycle. We plan to build on this work to extend the scope of

the meta-data collected on mappings, however in our approach the W3C PROV vocabulary will be used as the basis of lifecycle fields such as creation date.

A notable approach to publishing mappings however is the *R2R Framework* [8], which is a framework for executing mappings between RDF ontologies. The R2R framework has its own language called the *R2R mapping language*[9] for publishing mappings on the web. Similar to our approach of representing mappings, the R2R mappings are instances of a *Mapping* class, from the R2R mapping language, not the GLOBIC vocabulary. The mappings are modeled with meta-data and use the properties *sourcePattern* and *targetPattern* to represent the triple patterns which are executed by the R2R Framework. Our approach differs in that mapping instances are linked to a SPIN representation of a SPARQL construct query, which is ultimately executed by a SPARQL processor.

The SPARQL Inference Notation (SPIN) is a set of vocabularies that are used to represent business rules and constraints via SPARQL queries [9]. Tools that implement SPIN, such as TopBraid Composer[10] have been used in a wide range of applications [10-13]. Such tools also allow custom functions (which may not appear in the SPARQL specification) to be declared and executed, which make SPIN tools versatile at establishing data constraints and even data mapping [14].

5 Evaluation

This section describes two initial experiments of a series of planned experiments; the two here were carried out in order to evaluate our approach of representing mappings. The first experiment compared SPARQL's mapping capabilities with that of the R2R Framework. The second experiment involved testing the expressiveness of SPIN with regard to expressing SPARQL construct queries as RDF.

5.1 Experiment 1: Comparing SPARQL's mapping capabilities to the R2R Framework

Hypothesis: It is possible to represent all of the 70[11] R2R Framework test mappings as SPARQL construct queries and the execution of these SPARQL construct queries will produce identical results as the R2R Framework mapping results.
Method: First, a data-set[12] was collected. The creators of the R2R Framework devised 72 test mappings[13] between DBpedia and 11 other data-sources to test their framework. We collected instance data, related by the test mappings, via SPARQL endpoints and data dump files. Then the test mappings were executed against the data-set using the R2R Framework. This resulted in 70 output files consisting of new-

9 http://wifo5-03.informatik.uni-mannheim.de/bizer/r2r/spec/
10 http://www.topquadrant.com/tools/IDE-topbraid-composer-maestro-edition/
11 Note that only 70 of the 72 test mappings were carried out as data from BookMashup could not be obtained
12 Data from this experiment can be found at: http://www.scss.tcd.ie/~meehanal/Experiment1/
13 http://wifo5-03.informatik.uni-mannheim.de/bizer/r2r/examples/DBpediaToX.ttl

ly inferred triples. Next the data-set was loaded into an Apache Jena triple-store with Fuseki SPARQL server[14]. The 70 test mappings were represented as SPARQL construct queries and executed against the data in the triple-store. This resulted in 70 output files consisting of newly inferred triples. Lastly, the output files from the R2R Framework were compared with the respective output files derived from the SPARQL construct queries.

Results: It was found that the SPARQL construct queries created identical outputs as the R2R Framework for all of the 70 test mappings, which indicates that the SPARQL construct queries were accurate representations of the R2R Framework test mappings.

Discussion: The results are promising in showing that SPARQL is as capable as the R2R Framework for executing mappings on RDF data sets. The SPARQL 1.1 specification standardized a number of functions, such as string manipulation functions, which allow for more complex mappings to be carried out. Prior to SPARQL 1.1, the R2R Framework test mapping that involved string manipulation would not be possible using the SPARQL 1.0 specification, allowing the R2R Framework to represent more complex mappings than SPARQL.

5.2 Experiment 2: Testing SPIN's Expressivity

Hypothesis: It is possible to express all 70 of the SPARQL construct queries (from Experiment 1) as RDF via SPIN.

Method: This test used the SPIN RDF Converter and TopBraid Composer 4.4.0 (free edition) to transform the 70 SPARQL construct queries to SPIN syntax. An error is produced by the converter and composer if it cannot represent a SPARQL query.

Results: It was found that SPIN could represent all 70 of the SPARQL construct queries. The construct queries were categorized according to Scharffe's correspondences patterns [15]. All 70 fell into 3 patterns: *Equivalent Class*, *Equivalent Relation* and *Property Value Transformation* as illustrated in Figure 2 (some queries span across two patterns).

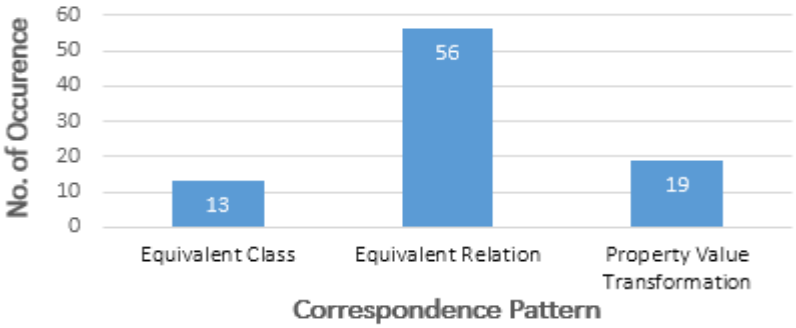


Figure 2. R2R Test Mappings broken down by Correspondence Pattern Type

[14] http://jena.apache.org/documentation/serving_data/

Discussion: Initially it was found that the SPIN RDF Converter could only represent 64 of the 70 SPARQL construct queries. Specifically the SPARQL 1.1 standardized functions *REPLACE*, *STRBEFORE* and *STRAFTER* could not be represented. The creators of SPIN were contacted and the SPIN RDF Converter is using an outdated version of SPIN and will be updated in the near future. However, TopBraid Composer 4.4.0 uses the latest version of SPIN and this was used to represent the 6 SPARQL construct queries that the SPIN RDF Converter could not.

6 Conclusions and Future Work

In this paper we have proposed a semantic mapping representation, between RDF data sources, that allows the mapping to be published, discovered and executed. The goal of the mapping representation is to provide an approach towards greater interoperability between heterogeneous tools operating within a localization tool-chain.

We represent the executable mapping statement as a SPARQL construct query which is expressed as RDF via SPIN. Mappings are modelled with meta-data, also expressed as RDF using the GLOBIC and W3C PROV vocabularies. All aspects of a mapping are published in a triple store alongside other data, where they can be discovered, queried and ultimately executed by a SPARQL processor.

We have shown that SPARQL construct queries are just as expressive as the R2R Mapping Language for representing a wide variety of mappings and that these queries can be represent in RDF via the SPIN syntax.

Future work will investigate a model of SPARQL-based mapping quality analytics and lifecycle management where all aspects of a mapping (meta-data and SPIN representation) can be queried and even updated/deleted using SPARQL queries.

Acknowledgements. This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of CNGL Centre for Global Intelligent Content (www.cngl.ie) at Trinity College Dublin.

References

1. Shvaiko, P. and Euzenat, J. "Ontology matching: state of the art and future challenges." *Knowledge and Data Engineering, IEEE Transactions on* 25, no. 1, 158-176, 2013.
2. Arch-int, N. and Arch-int, S. "Semantic Ontology Mapping for Interoperability of Learning Resource Systems using a rule-based reasoning approach." *Expert Systems with Applications* 40, no. 18, pp. 7428-7443, 2013.
3. Kumar, S. and Harding, J. A. "Ontology mapping using description logic and bridging axioms." *Computers in Industry* 64, no. 1, pp. 19-28, 2013.
4. Euzenat, J., Polleres, A. and Scharffe, F. "Processing ontology alignments with SPARQL." In *Complex, Intelligent and Software Intensive Systems, 2008. CISIS 2008. International Conference on*, pp. 913-917. IEEE, 2008.

5. Rivero, C. R., Hernández, I., Ruiz, D., and Corchuelo, R. "Generating SPARQL executable mappings to integrate ontologies." In *Conceptual Modeling–ER 2011*, pp. 118-131. Springer Berlin Heidelberg, 2011.
6. Keeney, J., Boran, A., Bedini, I., Matheus, C. J. and Patel-Schneider, P. F. "Approaches to Relating and Integrating Semantic Data from Heterogeneous Sources." In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp. 170-177. IEEE Computer Society, 2011.
7. Thomas, H., Brennan, R., and O'Sullivan, D. "Using the OM 2 R Meta-Data Model for Ontology Mapping Reuse for the Ontology Alignment Challenge–a Case Study." In *Proceedings of the 7th Intl. Workshop on Ontology Matching*, vol. 946. 2012.
8. Bizer, C. and Schultz, A. "The R2R Framework: Publishing and Discovering Mappings on the Web." In *1st International Workshop on Consuming Linked Data (COLD2010)*, Shanghai, China, November, 2010.
9. Knublauch, H. SPIN SPARQL Inferencing Notation. *http://spinrdf.org/* (accessed 06 03, 2014).
10. Fürber, C. and Hepp, M. "Using SPARQL and SPIN for data quality management on the Semantic Web." In *Business Information Systems*, pp. 35-46. Springer Berlin Heidelberg, 2010.
11. Spohr, D., Cimiano, P., McCrae, J. and O'Riain, S. "Using spin to formalise accounting regulations on the semantic web." In *International Workshop on Finance and Economics on the Semantic Web (FEOSW 2012)*, pp. 1-15. 2012.
12. Lefrançois, M. and Gandon, F. "ULiS: An Expert System on Linguistics to Support Multilingual Management of Interlingual Semantic Web Knowledge bases." In *MSW-Proc. 2nd Workshop on the Multilingual Semantic Web, collocated with ISWC-2011*, vol. 775, pp. 50-61. 2011.
13. Andreasik, J., Ciebiera, A. and Umpirowicz, A. "ControlSem–distributed decision support system based on semantic web technologies for the analysis of the medical procedures." In *Human System Interactions (HSI), 2010 3rd Conference on*. IEEE, 2010.
14. Kovalenko, O., Debruyne, C., Serral, E. and Biffl, S. "Evaluation of Technologies for Mapping Representation in Ontologies." In *On the Move to Meaningful Internet Systems: OTM 2013 Conferences*, pp. 564-571. Springer Berlin Heidelberg, 2013.
15. Scharffe, F. "Correspondence Patterns Representation." PhD thesis, University of Innsbruck, 2009.

WaSABi 2014: Breakout Brainstorming Session Summary

Sam Coppens[1], Karl Hammar[2], Magnus Knuth[3], Marco Neumann[4], Dominique Ritze[5], Miel Vander Sande[6]

[1] IBM Research - Smarter Cities Technology Center (SCTC), Ireland
[2] Information Engineering Group, Jönköping University, Sweden
[3] Hasso Plattner Institute, University of Potsdam, Germany
[4] KONA LLC, New York, USA
[5] Research Group Data and Web Science, University of Mannheim, Germany
[6] iMinds - Multimedia Lab, Ghent University, Belgium

1 Introduction

A key program point of the 2nd International Workshop on Semantic Web Enterprise Adoption and Best Practice (WaSABi 2014) was the breakout brainstorming session, at which challenges regarding enterprise adoption of Semantic Web technologies were discussed, and potential solutions to some of those challenges described.

The requested output of this session was for each group of 5 people to produce a few slides detailing such challenges and solutions, with an eye towards possible future collaboration areas between the academic and industry representatives.

Each group was provided with conversation starter sheets on enterprise adoption topics identified by the workshop chairs, and given an hour of time to discuss and summarise issues related to these topics. The resulting slides are attached to the WaSABi 2014 proceedings. Additionally, the contents of the presentations accompanying the slides are summarised in the following sections.

2 Challenges

One group discussed topics related to maintainability of systems based on Semantic Web technology. This is obviously a fundamental software quality issue that needs to be handled if enterprise is to adapt such technologies; systems must be maintainable for the duration of their lifetime, which may be significantly longer than the lifetimes that academics developing prototypes for research papers are accustomed to. Three challenges in particular were considered, though this list is far from exhaustive:

- *The longevity of URIs in a changing environment* - the distributed nature of the Semantic Web and the Linked Open Data cloud means that companies may make themselves dependant upon resources that are outside of their sphere of influence. What might happen if an entire critical namespace were to go offline due to lack of funding or interest? For instance, if

xmlns.com (which hosts the FOAF namespac) were to shut down, or purl.org, or schema.org, etc?
- *Persistence of vocabularies/datasets* - Related to the above, what might happen if a resource published on a particular namespace were to be modified in a way that is incompatible with existing systems that makes use of said resource?
- *Communication of popular technologies and vocabularies* - A lot of ontologies and technology are released by academia, but do not reach sufficient degrees of adoption to warrant maintaining in the long term. How can enterprises (and academics!) measure the adoption rate and subsequent chance of long-term survivability of such technologies and ontologies?

Another group discussed the issue of hiding complexity from software developers and simply making the tech work. For Semantic Web technology to be adopted, it is crucial that the learning curve be as flat as possible. Otherwise established methods and technologies, that may be poorer in a number of perspectives than their Semantic Web counterparts but are still sufficient to solve the task satisfactorily, will win out. Consequently, the Semantic Web community needs to consider how the complexity of the Semantic Web stack can be abstracted away and hidden from software developers. Here some issues were particularly singled out:

- *What are the reference solutions / best practices concerning some common and basic developer problems?* - Some examples where there are presently no single leading Semantic Web technology include: ORM/DAO transformation, ETL, and LDP/Middleware.
- *Where to make the cut skipping Semantic Web internals?* - That is, how much Semantic Web technology do we expose to the developer users, and how much do we abstract away? If we hide too much of the complex stuff, there is a risk that developers to not make full use of the power of Semantic Web technologies, but rather consider it "yet another data storage mechanism" or "just another API".
- *Scalability: Volume, Velocity, Variety* - There is in the research community still significant discussion on how to reach scalability using Semantic Web technologies, whether general purpose SPARQL endpoints are usable at all, whether REST-ful APIs are the way to go, whether Linked Data Fragments may be a way forward, etc. Developers do not want to deal with this uncertainty, nor having to make hard judgements about tradeoffs and technology ahead of getting started with a project; they just want tech that works sufficiently well and with the promise of future scalability.

3 Maintainability Challenges Solution Suggestions

Regarding the first set of problems, the maintainability of systems based on Semantic Web technology, a number of potential solutions were suggested. Again, this list of course not exhaustive, but may be a useful starting point.

3.1 The longevity of URIs in a changing environment

Developers building vocabularies that they expect to be around for some time, or to be used in a non-trivial context, should seriously consider under which namespace they publish that vocabulary. It is likely that using the web hosting provided by their current employer is a bad choice, as a simple change of jobs might break the namespace. The suggestion of the breakout group is to make use of the W3C Vocabulary Services[7], available to any W3C Community Group (i.e. you'll need at least five people accepting to be part of the community maintaining this namespace). The breakout group has a hard time seeing any organisation more likely to maintain a consistent and stable long-term namespace than the W3C.

When it comes to existing vocabularies that are hosted on namespaces that are not guaranteed to persist in the long term, the breakout group suggests that the vocabulary authors/hosts analyse the access logs of said vocabularies to ascertain the number of users that actually make use of these namespaces, before performing any changes to them. In the case that these resources are extensively used by the community (e.g., FOAF, schema.org, etc) it may be argued that such statistics be of public interest and ought to be released to the community if possible. This might also indicate adoption rates, which can help enterprises make technology choices.

Another idea that was brought up was studying whether the traditional DNS-based URIs are in fact necessary in the future. Considering that there are technologies such as BitTorrent, which uses DNS-less anchor links for resource resolution, possible similar technologies could be repurposed for the Semantic Web? The breakout group is not very knowledgeable about the internals of BitTorrent technology, but suggests that these types of technologies and solutions be studied also.

3.2 Persistence of vocabularies/datasets

The first recommendation of the breakout group concerning this topic was vocabularies and datasets, once they have been published on a publically resolvable URI, should **NOT** be changed. This is in line with the existing W3C Best Practice Recipes for Publishing RDF Vocabularies[8].

Additionally, the breakout group suggested that easy to understand and easy to interpret change logs between different versions of a particular ontology (published on different namespaces) would be very helpful, to let developers know whether to adapt their systems and dependant ontologies to newer versions of some ontology. As far as the breakout group knows, there is presently no standardised manner of displaying such deltas in a developer-friendly manner, and this would be a valued research contribution. Related to the same issue, there are a number of annotations in OWL supporting versioning, but these are only

[7] http://www.w3.org/2013/04/vocabs/
[8] http://www.w3.org/TR/swbp-vocab-pub/

sparingly used. The group recommends that vocabulary authors learn about and use these annotations to a greater degree.

A further suggestion for future applied research (possibly a MSc project?) was the development of methods and tools to analyze a code base (primarily in Java, as this is the dominant language for Semantic Web development) and produce reports on the technology and resource dependencies of that code base. Such a tool could help developers find dependencies such as vocabularies, software libraries, etc. that they might want to back up or clone into their own version management systems.

3.3 Communication of popular technologies and vocabularies

The breakout group suggests the use of LOV[9] and LODStats[10], which are concerned with exactly this type of work for vocabularies/ontologies.

Concerning technologies and solutions, no equivalent exists. One suggestion was to consider social networking ideas; possibly setting up a community site with a high degree of interaction among users, listing technologies, and providing information about the pros and cons of each, known installations, etc. Presently semantics are discussed on many different places on the web; GitHub, SemanticWeb.org, answers.semanticweb.com, the ONTOLOG community, etc. Perhaps integrating such content into one place, and combining this with a list of well known technologies and installations, could be a useful addition to the community.

4 Complexity Challenges Solution Suggestions

The group suggests that Semantic Web researchers need to become more familiar and comfortable with composing systems using abstraction layers that hide functionality. This necessitates a degree of quality assurance not historically associated with research software, as those abstractions layers, i.e. APIs, will need remain stable and supported over time. The Semantic Web research community has a lot to learn from our practitioner partners, with regards to testing, packaging, and maintaining software!

In terms of practical development, the group suggests employing REST-style architectures, which the practitioner web developer community are already familiar with. Standardising on this type of architecture also allows our software to more easily become interoperable with non-semantics based components and systems, to the benefit of both sides.

The group notes that research in Semantic Web technology is becoming more driven by practical applicability, as exemplified by technologies such as JSON-LD (supporting the use of semantics with existing toolsets), Turtle (supporting RDF that human developers can read), Linked Data Fragments (supporting less

[9] http://lov.okfn.org/
[10] http://stats.lod2.eu/

expressive but very scalable and efficient querying), and schema.org (supporting simple data content schemas that search engines understand). These technologies, and many others, trade a little bit of technical or scientific "purity" for everyday usability by web developers: in the opinion of the group a very worthy tradeoff to make. Some of these developments are being lead exclusively by companies; others by academics with an understanding of enterprise needs. The continuation of the WaSABi workshop series seems highly important for supporting both categories of developers.

Proceedings of the Second International Workshop on Finance and Economics on the Semantic Web (FEOSW 2014)

Held at the 11th Extended Semantic Web Conference (ESWC 2014)

May 26th, Anissaras, Crete, Greece

Preface FEOSW2014

Financial statements and their importance due the current economic situation make the headlines of the most important media all over the world on a daily basis. Beginning of July, 2014, one of the brightest stars of the Madrid stock market imploded after WiFi provider "Let´s Gowex" was forced to declare bankruptcy and admit that its chief executive and founder had falsified the company financial statements for at least the past four years.

According to the Financial Times, the board of Gowex said it had accepted the resignation of Jenaro García Martín after he took full responsibility for fake accounts. Gowex's collapse is likely to deal a heavy blow to investor confidence in the Spanish stock market and its regulators, coming so soon after a similar accounting scandal brought down Pescanova, the frozen fish group.

One of the major goals of the FEOSW workshop has always been enabling through technology a further Knowledge and Information Systems management of financial information. Semantic Technologies and Data Science are two powerful innovation waves, which will leverage the power of applying inference and intelligent querying to the financial domain, bridging the gap between old-fashioned human-oriented tedious analysis of financial data to a new generation of algorithms-based systems with full spectrum of financial markets coverage.

Finally, we hereby introduce the best works presented to the FEOSW workshop and we hope this could attain to bringing financial information management to their full potential. With that very purpose, we will commit on celebrating the next FEOSW generation in the context of the ESWC.

Ángel García-Crespo
Juan Miguel Gómez-Berbís
Mateusz Radzimski
José Luis Sánchez Cervantes

Organization

Organizing Committee

Angel García Crespo (University Carlos III of Madrid, Spain)
Juan Miguel Gómez Berbís (University Carlos III of Madrid, Spain)
Mateusz Radzimski (ATOS Research & Innovation, Spain)
José Luis Sánchez Cervantes (University Carlos III of Madrid, Spain)

Program Committee

Ricardo Colomo Palacios, PhD, University Carlos III of Madrid, Spain
Alejandro Rodríguez González, PhD, Universidad Politechnica de Madrid, Spain
Tomás Pariente Lobo, ATOS Research and Innovation, Spain
Miha Grčar, Jožef Stefan Institute, Slovenia
Achim Klein, Universität Hohenheim, Germany
André Freitas, Digital Enterprise Research Institute (DERI), Ireland
Gandhi Hernandez Chan, PhD, Universidad Tecnológica Metropolitana, Mexico
Yull Gallardo, University Carlos III of Madrid, Spain
Israel González Carrasco, PhD, University Carlos III of Madrid, Spain
José Luis López Cuadrado, PhD, University Carlos III of Madrid, Spain
Ioan Toma, PhD, Semantic Technology Institute (STI), Austria

Sponsorship

Ministry of Economy and Competitiveness under the project FLORA (TIN2011-27405)

University Carlos III of Madrid. SoftLab group.

Table of Contents

A Linked Data Approach to Sentiment and Emotion Analysis of Twitter in the Financial Domain

J. Fernando Sánchez-Rada[1], Marcos Torres[1], Carlos A. Iglesias[1], Roberto Maestre[2], and Esther Peinado[2]

[1] Universidad Politécnica de Madrid,
ETSI Telecomunicación, Avda. Complutense, 30, 28040 Madrid, Spain
[2] Paradigma Labs, Paradigma Tecnológico,
Avda. Europa, 26, Pozuelo de Alarcón, 28224 Madrid, Spain

Abstract. Sentiment analysis has recently gained popularity in the financial domain thanks to its capability to predict the stock market based on the wisdom of the crowds. Nevertheless, current sentiment indicators are still silos that cannot be combined to get better insight about the mood of different communities. In this article we propose a Linked Data approach for modelling sentiment and emotions about financial entities. We aim at integrating sentiment information from different communities or providers, and complements existing initiatives such as FIBO. The approach has been validated in the semantic annotation of tweets of several stocks in the Spanish stock market, including its sentiment information.

Keywords: linked data, semantic, finance, sentiment analysis, emotions

1 Introduction

The proliferation of user generated content in web sites and social networks, such as Facebook, TripAdvisor or Twitter, has lead to an increased awareness of the power of social networks for expressing opinions about products, services and even disasters. These so-called social sensors enable real time indexing of the social web with the aim of providing insight about the structure and activity of social networks. They provide a vast array of application possibilities, from monitoring brands or products to become early disaster warning systems [1].

In the financial field, social sensors can provide additional valuable information that complements other sources of information used in fundamental analysis, such as financial newspapers. In particular, sentiment analysis has been one of the most popular technologies to measure the investment mood. The sentiment stock market indicator has become a popular indicator that is provided together with the classical fundamental and technical stock market indicators [2]. Several websites provide the investor emotion index[3] or their sentiment, like AII Investor

[3] Market Emotion by CNN Money available at `http://money.cnn.com/data/fear-and-greed/`

Sentiment Survey[4], StockMarketSensor[5], or SentimentTrader[6], just to name a few.

In addition, recent research has shown that sentiment expressed in microblogging sites such as Twitter can be applied to predict daily changes in stock values [3,4].

Linked Data is another valuable resource that can provide financial analysts with an integration of available data sources in their activity [5]. Linked Data can provide a wide array of opportunities in the financial field. As reported by O'Riain et al. [6], depending on the information consumer needs, the integration and augmentation of financial information can lead to a significant benefit for financial and business analysis in tasks such as competitive analysis, fraud detection or figures comparison. It is also worth mentioning the recent trend towards open government and eGovernment data initiatives for public sector information, statistics data and economic indicators. The current status is promising, with a large volume of financial and economic data sets already available. Several researchers have shown this potential for different use cases, such as cross-lingual query of financial and business data from multiple sources [7,8], using social media in investment decisions [9,10] or enriching corporate financial reporting [11].

The aim of this article is the application of a Linked Data approach to expressing sentiments and emotions about financial concepts, which financial analysts can use to combine opinions expressed in different social media sites.

The article is arranged as follows. Sect. 2 gives an overview of the vocabularies we have defined for modelling sentiment and opinions as well as its interlinking with financial vocabularies such as FIBO [12]. Sect. 3 outlines our system design. Sect. 4 provides an overview of our experimental design and results. Sect. 5 expresses our conclusions and a brief discussion of future directions for this line of research.

2 Modeling Sentiment and Emotions as Linked Data

This section provides insight about the potential of Linked Data for accessing, interlinking and reasoning about business data sources. To leverage that power, it is necessary to have a robust representation model for sentiment in the financial context. Rather than creating an ad-hoc model, the Linked Data approach is to look for models for each domain and connect them. In particular, we will need a model for financial entities, a model for sentiment analysis results, and a model for microblogging messages. The following sections review the models (also referred to as ontologies or vocabularies) available in these domains, and Sect. 2.3 exemplifies the use of the final integrated model.

[4] AII Investor Sentiment Survey available at `http://www.aaii.com/sentimentsurvey`
[5] Available at `http://www.stockmarketsensor.com/`
[6] Available at `http://www.sentimentrader.com/`

2.1 Linked Data in the Financial Domain

Financial Industry Business Ontology (FIBO) [12] is a collaborative industry initiative to describe financial data standards using semantic technology. FIBO has been authored by Enterprise Data Management (EDM) council under the technical governance of the Object Management Group (OMG). FIBO has two distinct aspects: a business ontology and a presentation for business readability. FIBO is released in discrete ontologies by subject area: (i) Business Entities; (ii) Security, Loans, Derivatives and (iii) Corporate Actions and Transactions. At the time of this writing, only the first specification for Business Entities has been made public. The specification identifies a taxonomy of basic entities: Human Being, Legal Person, Organization and Legal Entity. This taxonomy is extended with other derived entities, such as Minor, Natural Person, Artificial Person (Company Limited by Guarantee, Legally Incorporated Partnership, Foundation or Incorporated Company), Formal Organization (Trust, Partnership or Incorporated Company) and Informal Organization. In addition, the ontology models concepts such as control and ownership.

Financial Exchange Framework Ontology (FEF) [13] is an ontology defined by International Financial Information Publishing (IFIP) Ltd. with the aim of providing an enterprise-wide publication and integration standard. FEF ontology provides support for modelling financial components and financial entities.

The FP7 FIRST Project (Large Scale Information Extraction and Integration Infrastructure for Supporting Financial Decision Making) has defined an ontology for sentiment analysis in financial domains [9,10]. The ontology identifies Orientation Term (OT), Financial Instrument (FI) and Indicator (I) and their relationships. In addition, the ontology conceptualises specialisations of FI (stocks and stock indexes), economic indicators, and relationships among them. Based on this ontology, the project FIRST has elaborated a set of ontologies for currencies, companies, financial instruments (stocks and stock indexes), funds, financial institutions, insurance companies and banks, available at FIRST project[7].

In its simple form, a FIBO definition would be a single triple. However, FIBO is a complete ontology that enables much more powerful assertions, as will be shown later.

2.2 Linked Opinions and Emotions about stocks

In this section we introduce two vocabularies, Marl and Onyx, that we have defined for providing a uniform vocabulary for expressing sentiments and emotions, respectively, according to linked data principles.

Marl [14] is a standardised data schema designed to annotate and describe subjective opinions expressed on the web or in particular Information Systems. Its aim is to show the benefits of publishing in the open, on the Web, the results of the opinion mining process in a structured form. On the road to achieving

[7] http://first.ijs.si/firstontology/

this, Marl attempts to answer the research question of to what extent opinion information can be formalised in a unified way.

Marl is the result of analysing the properties that characterise opinions expressed on the web or inside various IT systems. The final set of concepts proposed is shown in Fig. 1. It should be noted that opinions in Marl are meant to be linked to an entity. Such entity can be a FIBO Corporation, as described in the previous section. We will make use of this property in Section 2.3.

A detailed description of each particular property and an explanation of their meaning can be found in the vocabulary's specification [8].

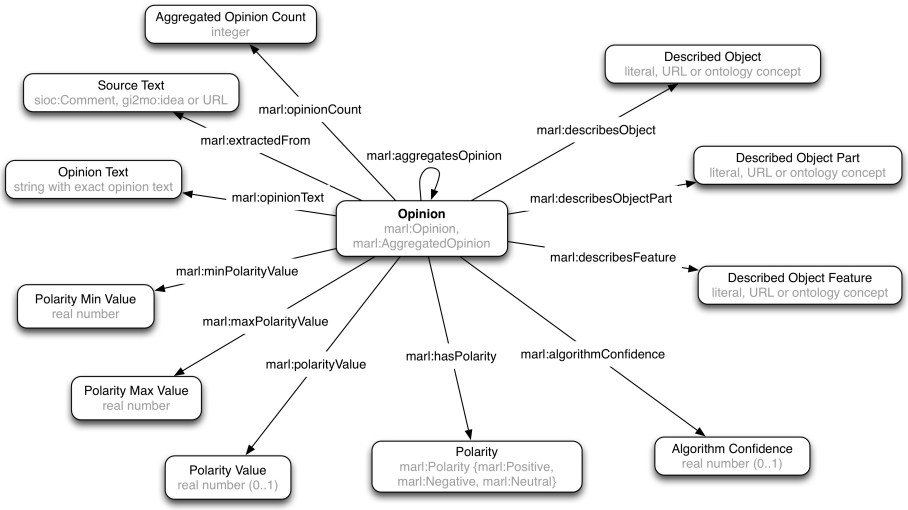


Fig. 1. Marl entities

Onyx [15] is a vocabulary to represent the Emotion Analysis process and its results, as well as annotating lexical resources for Emotion Analysis. It includes all the necessary classes and properties to provide structured and meaningful Emotion Analysis results, and to connect results from different providers and applications.

At its core, the Onyx ontology has three main classes: EmotionAnalysis, EmotionSet and Emotion. In a standard Emotion Analysis, these three classes are related as follows: an EmotionAnalysis is run on a source (generally in the form of text, e.g. a status update), the result is represented as one or more EmotionSet instances that contain one or more Emotion instances.

The specification of the Onyx vocabulary [9] contains an updated description of all its elements, with some usage examples.

[8] `http://www.gsi.dit.upm.es/ontologies/marl`
[9] `http://www.gsi.dit.upm.es/ontologies/onyx`

2.3 Using Linked Data

First of all, let us review a simplified version of the integration of all the elements that we described. To keep it as simple as possible, we will avoid any provenance information (such as who or how analised the twit to extract the opinion) or information about the post itself (author, date, etc.) This simplicity will not prevent us from harnessing the potential of Linked Data.

Listing 1.1. Simple representation using FIBO

```
ex:myOpinion a marl:Opinion;
              marl:hasPolarityValue marl:Positive;
              marl:describesObject ex:GSantander;
              marl:extractedFrom ex:twit1.
ex:twit1 a sioct:MicroblogPost;
         sioc:content "I like testing Grupo Santander".
ex:GSantander a fibo:IncorporatedCompany.
```

In this work, we have gathered thousands of posts from Twitter and stored them in a graph using a more complex version of this schema.

In order to provide a semantic representation of tweets, we have selected TwitLogic [16], which provides a vocabulary for tweets. The basic fields and their relationships are mainly RDF properties and classes taken from well-known sources like FOAF [17] or SIOC [18]. In this work we make use of FIBO to represent the entities of the financial domain. More specifically, we deal with Banks (Incorporated Companies) that have social presence and/or are mentioned by microblogging users. Marl and Onyx have been used for sentiment and emotion annotation, respectively. With this model, we can query all the opinions about a certain entity, statistics such as Positive/Negative ratio, and so on. Listing 1.3 shows an example that gets the count of positive and negative opinions about each entity.

However, the true potential of Linked Data comes into play when we use data from different sources. For instance, if there is another endpoint that contains opinions gathered from Twitter or other social networks, we can query their information seamlessly, provided they use Marl and FIBO as well.

If that example still seems uninteresting, we can also use disparate sources, such as DBpedia. DBpedia contains general information about many entities, which includes several corporations. To be able to query DBpedia, we just need to link our entities to a DBpedia entity. If we take our former example, this modification is as simple as:

Of course, this also involves named entity recognition techniques, which are covered in Section 3.2. Once this step is done, we can issue complex queries that answer questions such as: "What is the general opinion about Banks in Spain?", or "What is the relationship between year of incorporation and the number of opinions in social media?". Note that such queries could use advanced FIBO information, such as current contracts or date of incorporation.

Listing 1.2. Linking FIBO entities to DBpedia

```
ex:GSantander rdfs:seeAlso dbpedia:Santander_Group .
```

Listing 1.3. Query all positive opinions

```
PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX marl: <http://www.gsi.dit.upm.es/ontologies/marl/ns#>
SELECT ?entity
    COUNT(?negative_opinion) AS ?negative_opinions
    COUNT(?positive_opinion) AS ?positive_opinions
WHERE {
{
    ?positive_opinion marl:describesObject ?entity .
    ?positive_opinion marl:hasPolarity marl:Positive .
} UNION {
    ?negative_opinion marl:describesObject ?entity .
    ?negative_opinion marl:hasPolarity marl:Negative .
} } GROUP BY ?entity
```

3 Financial Twitter Tracker Architecture

In this section we describe the architecture of a prototype, called Financial Twitter Tracker, that we have developed for tracking the sentiment evolution of financial entities in Twitter. The core of the system is a semantic pipeline, described below, where tweets are retrieved and analysed. As a result, tweets are semantically annotated as stored in the semantic store Linked Media Framework (LMF) [19]. LMF also provides indexing capabilities based on Solr [20] full text indexing scalable solution. Finally a linked data visualisation framework called Sefarad[10] has been used in order to provide business analysts with a dashboard that assists them in their business decisions, as shown in Fig. 3.

The semantic pipeline for sentiment analysis consists of three tasks. First, the system connects to the Twitter API (Sect. 3.1) and retrieves tweets that match a list of predefined keywords. Then, a semantic analysis (Sect. 3.2 is carried out. Finally the sentiment analysis is done (Sect. 3.3).

3.1 Tweet retrieval

For the purpose of obtaining tweets we developed a wrapper over the services offered by the public Twitter API[11], concretely method search bounded by dates and keywords, which allows the retrieval of each and every tweet published within a particular day and regarding a particular topic. Given the data set of study, several related topics to financial world – such as banking, telecommunication, energy, to name a few – were established. Such data sets have been split according to different languages in order to increase performance and accuracy within the developed "sentiment analysis".

[10] Available at http://github.com/gsi-upm/Sefarad
[11] https://dev.twitter.com/docs/api/1.1/get/search/tweets

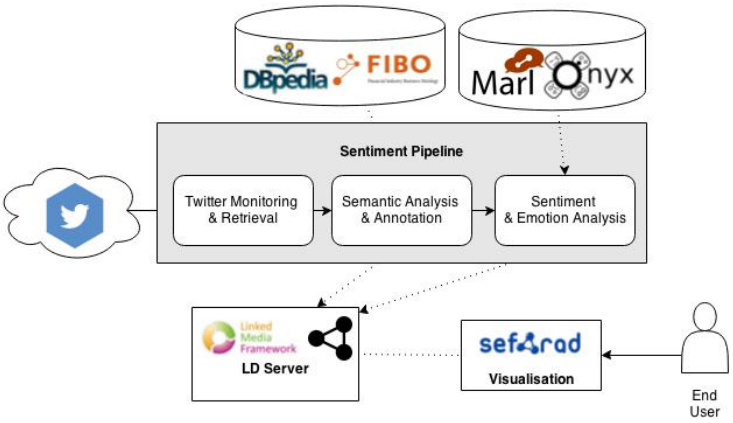


Fig. 2. Financial Twitter Tracker Architecture

3.2 Semantic Analysis and annotation

Data from Twitter is very heterogeneous, as it is used for different purposes (e.g. reviews, factual data, personal comments), covering different categories and subjects. Hence, it was necessary to carry out a categorization prior to the data analysis itself. With such filtering in mind, the Support Vector Machine system (SVM) was developed, taking into account the fact that it supports high dimensional data [21] and their suitability for classifying high volume of information using only support vectors which can be used in any distributed system [22,23,24] offering a great capability of cohesion and adaptation for the MapReduce paradigm. Several studies have proved that SVM provides better results than other techniques of classifications [25]. The system mentioned above has been trained throughout a random sampling of tweets tagged manually using Python with scikit [26] and numpy [27].

As POS-tagging, Treetagger [28] was chosen since it provides support for several languages. After acquiring a financial corpus for tracking a set of financial institutions, this corpus was cleaned, leaving aside irrelevant terms and stop words. Afterwards, collocations were extracted from the most frequent terms generating triplets with a structure domain-context-word (i.e. finance - profits - increasing). Once established these triplets, the following stage was to manually tag them by assigning a quantitative score to determine polarity and synset corresponding to WordNet 3.0 [29][10] basis. These triplets entitle the system to register texts providing scores thanks to the arrangements with WordNet, and leaning on MultiWordNet [30], WN-Affect [31], WN-Domains[32] and SentiWordNet[33]. For this goal, SentiWordNet has been extended in order to reasign scores for the finance domain. The method to enrich the lexicon stands out because its simplicity in terms of configuration, granting the chance of adding new languages easily or extending attached features (affects, domains, scores, etc.)

Another relevant aspect about the lexicon enrichment for its later storage and visualization was the extraction of entities by a NER based on Wikipedia, so that information is compared to the entities published by Wikipedia in order to work out the possible extraction from the text. Periodically the system brings the available information up to date with the new entries published on the online encyclopedia. Finally, that information is lined up with the financial ontology FIBO to provide data in a standardized way in accordance with the semantic web principles such as RDF/OWL, allowing the integration in other technical systems that adapts the given standard. Thanks to FIBO it is possible to provide a clear meaning - without ambiguities - for the financial terms.

3.3 Sentiment and Emotion Analysis

The last stage of the pipeline is in charge of the sentiment and emotion analysis. With a view to quantify the "sentiment" the procedure is to perform the arithmetic mean considering all the registered values recognized in the tweet and using simple rules like inverters (i.e. not). The emotion field can be extracted from the connection between triplets (aligned with WordNet 3.0) and WN-Affect. The outcome stems from the analysis of each tweet which was stored in a MongoDB NoSQL data base, which can handle high volume of information fulfilling the big data requirements of twitter processing.

3.4 Storage and visualisation

After the processing is done, all the triples are stored in an LMF instance, which provides SPARQL and Solr [20] endpoints. We built a generic visualisation framework, Sefarad, that uses these endpoints to display relevant information in any modern browser. This framework is modular and highly customisable. It already contains several plugins that use the power of D3 [12] to display the financial information in several ways. The plugins used, their configuration and location can be configured via an in-browser editor. One of its plugins allows the representation of public sentiment about each entity using Chernoff faces [34].

4 Experimentation

Throughout classification and Sentiment Analysis stages stages of the aforementioned pipeline, we performed experimentation with the obtained data. The classification step has been developed with an SVM trained for the recognition of two groups; finance and non-finance; which states whether the tweets are to continue to the next flow level or, on the contrary, are to be discarded.

Within Machine Learning there are two main discovery methods: supervised and unsupervised learning. In supervised learning, a series of manually tagged data are provided for the system training. On the unsupervised setting, it is the

[12] http://d3js.org/

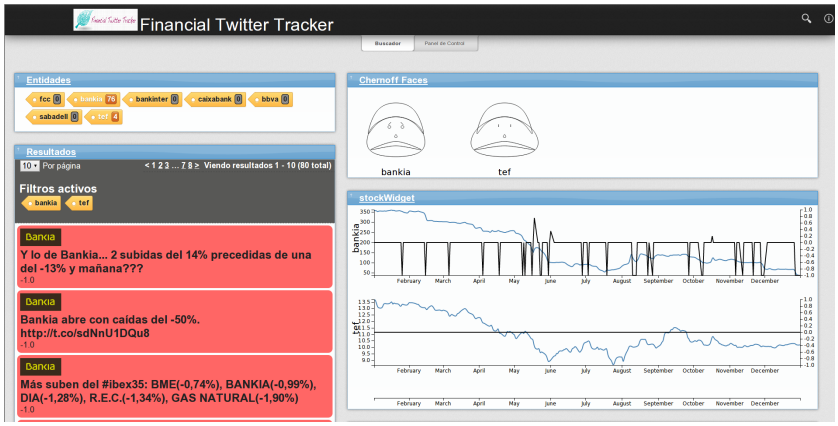


Fig. 3. Financial information displayed using Sefarad

system itself that directly infers patterns from the raw information. The current project uses supervised learning: a random set of tagged tweets has been trained by experts in finances and added to the established groups.

The model has been trained 4 times by modifying the range of information in order to measure and test the system. The first approach makes use of 90% of values for the training and 10% for the assessment, such proportions vary in the second training to 80%-20%, 70%-30% for the third, and 60%-40% for the last one [35]. Each of these cases has been tested five times in order to achieve the harmonic mean of the model accuracy with values chosen randomly for either experiment. The results of these experiments are summarised in Table 1.

Model training-Test	90%-10%	80%-20%	70%-30%	60%-40%
Average precision	0,940	0,9393	0,9369	0,9290
Supported Vectors	886,4548	825,4787	757,501	674,8602

Table 1. Results using different training options

From these results we observe that the bigger the quantity of information used to train the model, the more precise is the outcome, and that value decreases as the volume of data saved for the assessment grows. However, it is remarkable that the more data is used to train the system, the greater is the number of supporting vectors, and, consequently, the classifier loses computational performance.

The Sentiment Analysis phase has been tested against a set of manually annotated corpus. The evaluation was carried out in order to measure the effectiveness. Such accuracy conforms the ability the system owns to satisfy the feature it was developed for [36]. We have classified the results according to the parameters in Table 2. We used these values to define a set of quality metrics as shown in Table 2. The obtained results can be seen in Table 3.

System \ Expert	Identified	Not identified
Retrieved	a	b
Not retrieved	c	d

$$Recall := \frac{a}{a+c} \quad (1)$$

$$Precision := \frac{a}{a+b} \quad (2)$$

$$P_{omissions} := \frac{c}{a+c} \quad (3)$$

$$P_{falsepositive} := \frac{b}{b+d} \quad (4)$$

$$F_1 := 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

Table 2. Parameters used in the quality metrics.

Precision	Recall	Probability omissions	Probability false positive	F-Score
84'4%	63'87%	36'12%	77'04%	0.7271

Table 3. Resulting metrics.

5 Conclusions and Future Work

In this article we have presented a vocabulary for modelling sentiments and emotions. This vocabulary can be used to query opinions and emotions about financial institutions and stock values across different web sites and information sources. The main advantage of this approach is that heterogeneous sentiment indexes can be easily integrated and used together with other vocabularies such as FIBO. We have evaluated these vocabularies in a sentiment analysis service based on Twitter for tracking financial institutions.

As a future work, we are working on improving the visualisation and query capabilities of the interface so that non technical users, such as business analysts can take advantage of the possibilities that the Web of Data brings for exploring and consulting, sentiment about financial institutions in large amounts of complex and heterogeneous data.

Another current line of research is the standardisation of these vocabularies for sentiment and emotion. With this aim, we are participating in the Linked Data Models for Emotion and Sentiment Analysis W3C Community Group, which takes as a baseline the vocabularies Marl and Onyx.

6 Acknowledgement

This research has been partially funded by the Spanish Ministry of Industry, Tourism and Trade through the project Financial Twitter Tracker (TSI-090100-2011-114) and the EUROSENTIMENT FP7 Project (Grant Agreement no: 296277)

References

1. Chatfield, A.T., Brajawidagda, U.: Twitter early tsunami warning system: A case study in indonesia's natural disaster management. In: System Sciences (HICSS), 2013 46th Hawaii International Conference on. (Jan 2013) 2050–2060
2. Yardeni, E.: Stock market indicators: Fundamental, sentiment & technical. Technical report, Yardeni Research (2014) Available at `http://www.yardeni.com/pub/peacockbullbear.pdf`.
3. Vu, T.T., Chang, S., Ha, Q.T., Collier, N.: An experiment in integrating sentiment features for tech stock prediction in twitter. In: 24th International Conference on Computational Linguistics. (2012) 23
4. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science **2**(1) (2011) 1–8
5. O'Riain, S., Curry, E., Harth, A.: XBRL and open data for global financial ecosystems: a linked data approach. International Journal of Accounting Information Systems **13**(2) (June 2012) 141–162
6. O'Riain, S., Harth, A., Curry, E.: Linked Data Driven Information Systems as an Enabler for Integrating Financial Data. In: Information Systems for Global Financial Markets: Emerging Developments and Effects. IGI Global (2012)
7. Krieger, H., Declerck, T., Nedunchezhian, A.: MFO - the federated financial ontology for the monnet project. In: Proceedings of the 4th International Conference on Knowledge Engineering and Ontology Development, Barcelona, Spain (2012)
8. O'Riain, S., Coughlan, B., Buitelaar, P., Declerk, T., Krieger, U., Marie-Thomas, S.: Cross-lingual querying and comparison of linked financial and business data. In: Proceedings of 10th Extended Semantic Web Conference (ESWC), Montpellier, France (2013)
9. Grcar, M., Häusser, T., Ressel, D.: D3.1 semantic resources and data acquisition. Technical report, First project (2011)
10. Klein, A., Altuntas, O., Häusser, T., Kessler, W.: Extracting investor sentiment from weblog texts: A knowledge based approach. In: Proc. of the 2011 IEEE Conference on Commerce and Enterprise Computing. (2011) 1–9
11. Goto, M., Hu, B., Naseer, A., Vandenbusshe, P.I.: Linked data for financial reporting. In: 4th International Worshop on Consuming Linked Data (COLD2013), CEUR Workshop proceedings (2013)
12. Council, E.: FIBO. Financial Industry Business Ontology. Available at http://www.edmcouncil.org/financialbusiness (June 2013)
13. IFIP: FEF. financial exchange framework ontology. Available at http://www.financial-format.com/index.html (June 2003)
14. Westerski, A., Iglesias, C.A., Tapia, F.: Linked Opinions: Describing Sentiments on the Structured Web of Data. In: Proceedings of the 4th International Workshop Social Data on the Web. (2011)
15. Sánchez-Rada, J.F., Iglesias, C.A.: Onyx: Describing Emotions on the Web of Data. In: Proceedings of the 1st International Workshop on Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI, Torino, Italy, AI*IA, Italian Association for Artificial Intelligence (December 2013)
16. Shinavier, J.: Real-time# semanticweb in $\leq$ 140 chars. In: Proceedings of the Third Workshop on Linked Data on the Web (LDOW2010) at WWW2010. (2010)
17. Golbeck, J., Rothstein, M.: Linking social networks on the web with foaf: A semantic web case study. In: AAAI. Volume 8. (2008) 1138–1143

18. Breslin, J.G., Decker, S., Harth, A., Bojars, U.: Sioc: an approach to connect web-based communities. International Journal of Web Based Communities **2**(2) (2006) 133–142
19. Kurz, T., Schaffert, S., Burger, T.: Lmf: A framework for linked media. In: Multimedia on the Web (MMWeb), 2011 Workshop on, IEEE (2011) 16–20
20. Smiley, D., Pugh, D.E.: Apache Solr 3 Enterprise Search Serve. Packt Publishing (2011)
21. Dilrukshi, I., De Zoysa, K., Caldera, A.: Twitter news classification using svm. In: Computer Science & Education (ICCSE), 2013 8th International Conference on, IEEE (2013) 287–291
22. Jakkula, V.: Tutorial on support vector machine (svm). School of EECS, Washington State University (2006)
23. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREC. (2010)
24. Balahur, A., Steinberger, R., Goot, E.v.d., Pouliquen, B., Kabadjov, M.: Opinion mining on newspaper quotations. In: Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on. Volume 3., IET (2009) 523–526
25. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics (2002) 79–86
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12** (2011) 2825–2830
27. Oliphant, T.E.: Guide to NumPy, Provo, UT. (March 2006)
28. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of international conference on new methods in language processing. Volume 12., Manchester, UK (1994) 44–49
29. Fellbaum, C.: WordNet. Wiley Online Library (1999)
30. Pianta, E., Bentivogli, L., Girardi, C.: Multiwordnet. developing an aligned multilingual database. In: Proc. 1st International Conference on Global WordNet. (2002)
31. Strapparava, C., Valitutti, A.: Wordnet affect: an affective extension of wordnet. In: LREC. Volume 4. (2004) 1083–1086
32. Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A.: The role of domain information in word sense disambiguation. Natural Language Engineering **8**(4) (2002) 359–373
33. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: LREC. Volume 10. (2010) 2200–2204
34. Chernoff, H.: The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association **68**(342) (1973) 361–368
35. Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI. Volume 14. (1995) 1137–1145
36. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Volume 1. Cambridge university press Cambridge (2008)

Analyzing Stock Market Fraud Cases Using a Linguistics-Based Text Mining Approach

Mohamed Zaki[1,*] and Babis Theodoulidis[2]

[1] Cambridge Service Alliance, Department of Engineering, University of Cambridge, Cambridge, UK
mehyz2@cam.ac.uk
[2] Manchester Business School, University of Manchester, Manchester, UK
b.theodoulidis@manchester.ac.uk

ABSTRACT. The paper proposes a linguistics-based text mining approach to demonstrate the process of extracting financial concepts from the Security Exchange Commission (SEC) litigation releases (LR). The proposed approach presents the extracted information as a knowledge base to be used in market monitoring surveillance systems. Also, it facilitates users' acquisition, maintenance and access to financial fraud knowledge and improves search results in the SEC enforcement portal. Answering questions such as: who are the agents involved in the manipulation? Which patterns are associated with this manipulation? When was this manipulative action performed? This paper used the financial ontology for fraud purposes introduced by [19] to provide underlying framework for the extraction process and capture financial fraud concepts from the SEC-LR. In particular, text mining analysers have been developed to extract metadata concepts (e.g. 'LR No.', 'dates') and stock market fraud concepts (e.g. agents and manipulation types) from the actual SEC fraud case.

Keywords

Text mining, stock market fraud, stock market fraud ontology, stock market monitoring and surveillance system.

1 Introduction

The finance domain has suffered from a lack of efficiency in managing vast amounts of financial data, a lack of communication and knowledge sharing between analysts. Particularly, with the growth of fraud in financial market, cases are challenging, complex, and involve huge information that needs to be analyzed similarly to other legal cases. Gathering facts and evidence from the information available is often a very complex process. The impetus to effectively and systematically address stock financial market efficiency, including factors such as stock price manipulation, has long presented a very dynamic challenge to academia, the industry and relevant authorities. Interestingly, since 1960 the Security Exchange Commission (SEC) has prosecuted more than 24,525 fraud cases.

Many authors [1,10,12,15] have tested the impact of non-competitive behaviour in the stock market, and verified the possibilities of market manipulation. [1] presented a theoretical framework for profitable market manipulations, and provided empirical evidence using a comprehensive dataset of manipulation cases which occurred in the US stock markets and were published in SEC litigation releases from 1990 to 2001. Manipulators can artificially increase securities prices and make profits using various strategies, from classic manipulative trading practices that influence prices, to the sophistication of spam and scam manipulation using various internet channels [1]. Since the passing of the Securities Act in 1930, there is evidence that market manipulation has a significant impact on the efficiency of the securities market [11].

Despite the existence of many authoritarian regulations such as the Securities Exchange Act of 1934 and European Market Abuse Directive 200, prosecutors find it challenging to prepare a case with appropriate evidence, and the gain for an exchange of a successful prosecution would be small in comparison to the efforts and resources necessary to bring a criminal case. Fraud cases can be extremely complex and difficult to demonstrate to juries. That is why regulators only select certain cases for prosecution and prioritize instances of organized manipulation. Furthermore, few courts have experience in trying securities fraud cases.

There is an urgent need to develop a novel approach that could help regulators and relevant authorities in managing vast quantities of financial data, providing better communication and knowledge sharing among analysts, providing a mechanism to demonstrate knowledge of the processes of financial fraud, understanding and sharing financial fraud logic operations, managing relevant facts gathered for case investigations, and allowing reuse of these knowledge resources in different financial contexts [8].

Therefore, the continuous improvement and development of financial market monitoring and surveillance systems with high analytical capabilities to capture the fraud is essential to guarantee and preserve an efficient market [14]. Currently, these systems are used in limited fashion and act as a reporting archive for multiple functions within the exchange. Thus, this paper aims to provide significant cross-fertilization between financial research studies and information technology as it attempts to incorporate text mining techniques for the analysis as one of the most appropriate technological area, allowing analysis of stock market fraud documents through the development of linguistic and non-linguistic patterns. In this context, text mining is used to extract financial concepts from the SEC litigation releases and thus, provide an appropriate knowledge base about financial market manipulations. The paper provides empirical evidence of how text mining could help the market monitoring surveillance systems to explore the potential efficiency and effectiveness benefits for analysing litigation releases.

2 Previous Work

This section provides a review on existing market monitoring surveillance systems and fraud detection studies that used data mining and text mining to investigate and

detect fraudulent behaviors and ascertaining evidence of potential cases of fraud within different financial markets. In fact, these systems could support financial organizations to proactively detect transactions where market abuse is suspected.

Focusing specifically on the market manipulations domain, [9] described how the National Association of Securities Dealers (NASD) Inc. used a fraud detection system [called Advanced Detection System (ADS)] to monitor trades, to detect and identify any suspicious trading behavior for further investigation in the NASDAQ stock market. [6] work describe the Securities Observation, News, Analysis and Regulation system (SONAR), also developed by NASD. The system's main purpose is monitoring NASDAQ transactions in the stock market to detect and identify any potential insider trading and any falsification of news stories for the purposes of fraud. The work of [13] introduced a monitoring system used on the Thai Bond Market, which was commissioned by the Thai Bond Market Association (ThaiBMA). The market uses a real time approach to monitor transactions, investigate any unusual ones, and notify regulators where enforcement action is required. [4] proposed a market monitoring framework, comprising of the analysis components, tasks and flows of information of a complete financial market monitoring system. The framework is designed to have a past time or reactive monitoring engine, which is fed with either structured or unstructured data sources.

Many studies show how data mining and text mining techniques can be used in such domain. For example, [5] utilized data mining techniques (C4.5, decision tree, neural network, K-mean clustering, and logistic regressions) for the early detection of insider trading manipulation schemes before the news broke within the option market. [16] generated a conceptual framework to identify the individuals (and their communities) involved in trade-based manipulation using data mining such as Euclidian Distance (ED), Shared Nearest Neighbour (SNN), density-based algorithm (DBSCAN) and graph-partitioning algorithm (METIS). [2] employed a data mining approach to analyze two cases of manipulation in the New York Stock Exchange. The researcher used the decision trees technique to distinguish between manipulations and normal trading and to improve organizational fraud detection systems. [18] described a case study on fraud detection using data mining techniques that help analysts to identify possible instances of touting based on spam emails in the Pink Sheets market. Various data mining techniques such as decision trees, neural networks and linear regression are utilized in this emerging domain. [17] presents an exemplar case study of text mining and data mining to analyze the impact of 'stock-touting' spam e-mails and misleading press releases on trading data a real case from the over-the-counter (OTC) market, and which was prosecuted by the SEC. [3] presents a high frequency trading analysis of a particular trading scenario and discusses how quote stuffing can affect the function of trading systems.

This research contributes to the development of a comprehensive domain ontology for stock markets. Currently, the existing market monitoring systems lack a comprehensive financial knowledge base [19]. This research contributes to the development by providing an additional context for the evaluation of the domain ontology and furthermore, demonstrates how data sources such as the SEC litigation releases could be analysed using a text mining approach and could support fraud analysts in the in-

vestigation process. Finally, the paper uses SEC cases as the data source that has not been addressed in previous work.

3 Methodology

This paper used the financial ontology for fraud purposes introduced by [19] to provide underlying framework for the extraction process and capture financial fraud concepts from the SEC litigation releases. The ontology has a comprehensive financial concept system for fraud purposes, which utilized to semantically rich the knowledge base of market monitoring surveillance systems to potentially help fraud analysts to understand different manipulation patterns from prosecuted cases. In this context, this paper evaluates this ontology through a specific text mining instantiation that demonstrates the published prosecuted case in an appropriate fraud knowledge base. The role of the analysers is to identify the knowledge that lies in the prosecuted cases to be able to answer questions similar to those asked by users and analysts reading the cases themselves. Some key questions that the text mining analysers should answer include: Who is (are) the agent(s) involved in the manipulation? Which asset is being targeted? In which venue is the manipulation taking place? Which action has been performed or is planned? Which patterns are associated with this manipulation? When was this manipulative action performed? Where is the manipulator getting his profit?

3.1 Data Source

This paper used the SEC published litigation releases which are prosecuted fraud cases for the US stock markets as a main data source of fraud knowledge to be analyzed. The litigation releases are concerning civil lawsuits brought by the Commission in federal court. Each litigation release has a release number, release publication date, and action that include the defendants' names; and most of the releases have an external link to SEC complaint.

The SEC complaint documents are the reports produced by the US district court that describe violations cases of securities laws. In these documents the court describes in detail all the evidence and facts that make its decisions to prohibit the acts or practices that were the results of violation of the law or commission rules. The SEC complaint document structure consists of five main sections: 'Civil case action no' which is the file number the court allocates to the document; 'district court name', 'court clerk's office stamp' which includes his name, signature, title and filing date, 'title' including the names of the defendant and the plaintiff, and the main 'document sections' which generally include a summary of the complaint, list of defendants and relevant person entities involved in the violation, jurisdiction and the venue of the court, facts that describe the manipulation scheme and all the evidence that has been collected by the Commission to prosecute the defendants, 'fraud for reliefs and violation' includes all the acts and laws that defendants violated in such case.

Different textual sources were collected from the SEC website, such as RSS format, HTML and PDF files. The RSS format is used to download the recent litigation releases published in the SEC website. Most of the litigation releases have an HTML link that contains a short description as a summary of the case followed by a detailed description of the cases (SEC Complaints). The SEC complaints are PDF documents produced by the district courts to provide a full description of the prosecuted cases. Based on the case, the average size of these documents could vary from 10 to 60 pages. In particular, the text-mining application analyzed the third quarter of 2012 that contains 62 litigation releases with total size of the corpus is 185.1 MB.

3.2 Text Mining Application Design

This section demonstrates the design of text mining, which is constructed for the SEC fraud cases. As shown in figure 1, this study adapted the information extraction application layer introduced by the stock market fraud ontology [19]. The application layer can process all document types such as text, audio and video. Thus the TextualFormat class will have concepts related to documents used in the application, such as the SEC case study, the SEC litigation release and RSS feed. In addition, each document instance could be related to one or more annotations used to develop the linguistic patterns. Two main types of resource are used in the text-mining components, namely language resources and processing resources. Language resources contain resources such as a thesaurus, list of terms, concepts, synonyms, and types (semantic groupings of concepts). Processing resources incorporate analyzers, generators, recognizers (e.g. speech transcribers, handwriting recognizers), and retrievers (e.g. search engines).

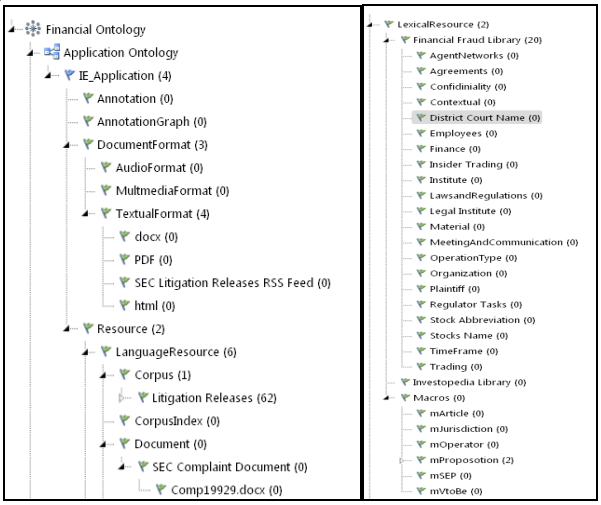


Fig 1 Text Mining Application Layer [19]

The 'corpus' concept has the actual litigation releases of the third quarter of 2012 that have been analyzed in the application, such as 'LR-22420' and 'LR-22421'. The

'document' class contains the 'insider trading' case study of Bio-Medicus, Inc.[1], which was originally a PDF file but has been converted to the 'docx' extension. This case will be used to demonstrate how text mining could automate the process of classifying financial concepts in the proposed classes in the financial fraud ontology.

The developed text-mining application uses different components to develop the advanced linguistic patterns, which could contain the following components: sub-classes from the developed library, synonyms, macros, and word gaps. 'FinancialFraud Library' is another sub-class added to the 'lexicalResource' concept which has 20 classes.This library was developed to help the text-mining application to extract concepts from the litigation releases, especially the fraud-related concepts. Based on the [19], the library has over 223 concepts that are classified and mapped to classes.

This library is used to develop and construct the advanced linguistic patterns. For example, "Confidentiality" includes a list of terms related to confidential information such as "confidential information, confidentiality, confidential advice, confidentiality agreement, confidentiality policy, code of ethical conduct, confidential, etc". "OperationType" contains terms related to trading operation such as "acquired, sold, obtained recommended, sell, buy, purchase, etc". "AgentNetwork" includes all terms representing the type of relatives who help manipulators to execute the fraud, such as "son, cousin, friend's wife, friend, relative, etc". "Employees" has all terms related to employees such as "employee, manager, chairman, board, office manager". "District Court Name" holds a list of different state courts such as "federal district court, eastern district of New York, middle district of Florida, district of New Jersey".

Furthermore, the library includes synonyms, which associate two or more concepts that have the same meaning. In particular, synonyms have been used to resolve the issue of misspelled concepts, and concepts having the same meaning, e.g. "Securities and Exchange Commission" and "Commission".

The macros is another class added to the lexicalResource concept, which represents reusable patterns; it is used to simplify the appearance of literals and word strings needing to be extracted, e.g. prepositions, articles, and verbs. Overall, the application includes six macros that support the extraction process and pattern development. For example, "mPreposition" includes a list of tokens related to prepositions such as "to, from, for, of, on, at, with, about, into, etc". "mArticles" includes a list of tokens related to English language articles such as "a, the, an, etc". "mVtobe" is another macro that holds tokens related to concepts concerning the verb 'to be, e.g. "is, are, was, were", , etc.

Process resource concept demonstrates the text-mining process and the advanced linguistic pattern approach employed on the basis of natural language processing (NLP) in order to linguistically analyze the litigation releases. 60 advanced linguistic patterns were developed to analyze the litigation releases sentence-by-sentence and to apply focus group participants' recommendations. This section demonstrates the 'Insider Trade' analyzers developed to automate the process of extracting information in the context recommended by the ontology to explain the fraud cases

[1] Case web link at http://www.sec.gov/litigation/complaints/2006/comp19929.pdf

4 Text Mining Analysis

This paper used the IBM-SPSS Modeler14 data and text mining workbench to develop the text mining analyzers [7]. The text mining application contains two components metadata analysers and SEC Complain Document Analysers.

4.1 Metadata Analysers

The text-mining metadata analysers aim to extract key metadata information from the data source. The target concepts are 'litigation release number', 'release publication dates', 'agents' which are the defendants' (individual or organization) names, 'document format type', the 'document link', 'civil case no.' which is the allocation number issued by the court, 'district court no.' including state courts or federal courts, and the 'plaintiff'. The text-mining application used some of the predefined libraries incorporated in the IBM PASW 14 software, such as date, time, person, organization. However, these patterns did not capture all related concepts within the document. Thus, extra patterns have been developed to cover the gap and to increase the accuracy of extraction. Furthermore, other analyzers were developed from scratch to match the specific structure of linguistic patterns.

Table 1 explains the patterns developed by the 'Civil Case No' analyzer. Each court has a different pattern and in order to capture most of them, 18 linguistic patterns using regular expressions were developed. These patterns have been numbered sequentially from 1–18 to avoid any break in numbering which might cause suspension or conflict when processing the document. For example, in the first pattern in Table 1, "6-12-CV-00932-JA-GJK", the regular expression was developed as [0-9]{1, 2} to match a digit repeated exactly one or two times [e.g. 6], followed by specific character "–", similarly, s[0-9]{1, 2} to match the two digits, followed by "-", followed by two character [case sensitivity is considered] [a-zA-Z] {1,2}, followed by five numbers [0-9]{3, 5}, followed by 2 characters [a-zA-Z]{1,2}, followed by three character [a-zA-Z]{1,3}. The "Civil Case No" analyzer successfully captured 100% from the 62 litigation releases used in the application.

From the 62 litigation releases, the analyzers' 'litigation release number', 'release publication dates', 'actions', 'document format type', 'document link', 'short description', 'detailed description', and 'plaintiff' have a high accuracy level of almost 98 % precision. Regarding 'Civil case no.' the analyzer achieved 98.93% precision, and for 'District court Name' 93.55% precision. In five releases the court names were not included due to the pending status of the allegation or administrative proceeding status, or were missed.

Table 1. 'Civil Case No' Analyzer

Item	Regular Expression developed Patterns	Examples of 'Civil case no.' Patterns
1	regexp1=[0-9]{1,2}-[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{3,5}-[a-zA-Z]{1,2}-[a-zA-Z]{1,3}	Civil Action No. 6-12-CV-00932-JA-GJK
2	regexp2=[0-9]{1,2}:[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{3,5}-[a-zA-Z]{1,3}-[a-zA-Z]{1,3}	Case No. 2:12-cv-03794-JLL-MAH
3	regexp3=[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{3,5}-[a-zA-Z]{1,3}-[a-zA-Z]{1,3}	Case No. 09-cv-7594-KBF-THK
4	regexp4=[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{4,5}-[a-zA-Z]{1,3}	Civil Action No.07-CV-1643-D
5	regexp5=[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{4,5}	Case No.12-CV-6421
6	regexp6=[0-9]{1,2}:[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{3,5}-[a-zA-Z]{1,3}	Civil Action No. 1:12-CV-02831-ODE
7	regexp7=[0-9]{1,2}:[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{3,5}	Civil Action No. 3:12-CV-519
8	regexp8=[0-9]{1,2} [a-zA-Z]{1,3}. [0-9]{4,5}	Civil Action No. 12 Civ. 5751
9	regexp9=[a-zA-Z]{1,2}-[0-9]{1,2}-[0-9]{1,4} [a-zA-Z]{1,3}	Civil Action No. CV-11-0137 WHA
10	regexp10=[a-zA-Z]{1,2}[0-9]{1,2}-[0-9]{1,4} [a-zA-Z]{1,3}	Civil Action No. CV12-1179 Jst
11	regexp11=[a-zA-Z]{1,2} [0-9]{1,2}-[0-9]{1,4} [a-zA-Z]{1,3}	Civil Action No. CV 10-8383 DSF
12	regexp12=[a-zA-Z]{1,4}-[0-9]{1,6}[a-zA-Z]{1,3}-\([a-zA-Z]{1,4}\)	Civil Action No.SACV-1213271JST-(JPRX)
13	regexp13=[a-zA-Z]{1,2} [0-9]{1,2}-[0-9]{1,5}	Civil Action No. CV 12-3200/cv 11-09859
14	regexp14=[0-9]{1,2}-[a-zA-Z]{1,3}.?-[0-9]{3,4}	Civil Action No.12-CIV-5100/12-CIV.-5550
15	regexp15=[0-9]{1,2} [a-zA-Z]{1,2} [0-9]{3,4}	Civil Action No.07 CV 3444
16	regexp16=[0-9]{1,2}-[0-9]{1,2}-[0-9]{1,3}	Civil Action No. 1:09-361
17	regexp17=[0-9]{1,2} [0-9]{1,4}	Civil Action No. 08 2457
18	regexp18=[0-9]{1,2}-[0-9]{1,3}	Civil Action No. 12-134

Despite the size of the documents, the analyzers are good enough to demonstrate how text mining could be used for automating the analysis of SEC litigation releases. Table 2 shows the final output of the target concepts captured by the data source analyzers. For example, in litigation release number 'LR-22396' published on 20th June 2012, the defendants are 'Gary J.Mortal', 'Martel Financial Group', and 'MFG Funding'. The user can find short or detailed descriptions of the release via the link http://www.sec.gov/litigation/litreleases/2012/lr22396.htm. Therefore, the only information provided by the SEC is through the link as the release does not yet have a complain file (See also has null value) issued by the respective court. The Security and Exchange Commission, the plaintiff in this release, sent the case to the federal district court. The civil case number of the release is '12-cv-11095'.

Table 2 Metadata Data Source Ontology Text Mining Analyzers' Output

	Release No.	Date	Action (Defendants)	Short Description	Detailed Description
1	LR-22396	Wed, 20 Jun 2012 16:03:29 EDT	Gary J. Martel, d/b/a Martel Finan...	SEC CHARGES MASSACHUSETT...	Gary J. Martel, d/b/a Martel Fi...
2	LR-22398	Mon, 25 Jun 2012 13:59:04 EDT	Ralph R. Cioffi and Matthew M. T...	Court Approves SEC Settlements ...	Ralph R. Cioffi and Matthew ...
3	LR-22399	Mon, 25 Jun 2012 16:26:34 EDT	Gurudeo Persaud	The Securities and Exchange Co...	Gurudeo Persaud: Lit. Rel. N...
4	LR-22400	Mon, 25 Jun 2012 16:26:34 EDT	Manuel M. Bello, Ayuda Equity F...	The Securities and Exchange Co...	Manuel M. Bello, Ayuda Equity...
5	LR-22401	Wed, 27 Jun 2012 10:43:55 EDT	Tai Nguyen	SEC Charges Founder of Equity R...	Tai Nguyen | U.S. Securities ...
6	LR-22402	Wed, 27 Jun 2012 10:43:55 EDT	AMMB Consultant Sendirian Ber...	SEC SUES FUND ADVISER FOR ...	AMMB Consultant Sendirian ...
7	LR-22403	Thu, 28 Jun 2012 12:41:52 EDT	Harbinger Capital Partners LLC ...	SEC Charges Philip A. Falcone a...	Harbinger Capital Partners L...
8	LR-22404	Thu, 28 Jun 2012 13:27:09 EDT	H. Clayton Peterson et al.	SEC Obtains Final Judgments On...	H. Clayton Peterson et al. | ...
9	LR-22405	Thu, 28 Jun 2012 13:27:09 EDT	FalconStor Software, Inc.	FalconStor Software, Inc.	FalconStor Software, Inc. | U...

	Release No.	Document Link	See Also	Document Type	Civil Case No.	DistrictCourtName	Plaintiff
1	LR-22396	http://www.sec.gov/litigation/litreleases/2012/lr22396.htm	$null$	htm	12-cv-11095	federal district court	securities and exchange commission
2	LR-22398	http://www.sec.gov/litigation/litreleases/2012/lr22398.htm	$null$	htm	08 2457	eastern district of new york	securities and exchange commission
3	LR-22399	http://www.sec.gov/litigation/litreleases/2012/lr22399.htm	$null$	htm	6-12-cv-00932-ja-gjk	middle district of florida	securities and exchange commission
4	LR-22400	http://www.sec.gov/litigation/litreleases/2012/lr22400.htm	$null$	htm	2:12-cv-03794-jll-mah	district of new jersey	securities and exchange commission
5	LR-22401	http://www.sec.gov/litigation/litreleases/2012/lr22401.htm	$null$	htm	12-cv-5009	southern district of new york	securities and exchange commission
6	LR-22402	http://www.sec.gov/litigation/litreleases/2012/lr22402.htm	$null$	htm	1:12-cv-01052	district of columbia	securities and exchange commission
7	LR-22403	http://www.sec.gov/litigation/litreleases/2012/lr22403.htm	$null$	htm	12-cv-5029	southern district of new york	securities and exchange commission
8	LR-22404	http://www.sec.gov/litigation/litreleases/2012/lr22404.htm	$null$	htm	11 civ. 5448	southern district of new york	securities and exchange commission
9	LR-22405	http://www.sec.gov/litigation/litreleases/2012/lr22405.htm	$null$	htm	cv 12-3200	eastern district of new york	securities and exchange commission

4.2 SEC Complain Document Analysers

The text-mining application analyzed the SEC complaint document produced by the US district courts. 11 analyzers have been developed to extract the annotated financial concepts. The developed analyzers analyzed the document sentence-by-sentence. In total, 60 advanced linguistic patterns were developed to extract infor-

mation related to financial fraud and classify this information in the appropriate ontology classes, as guided by the financial ontology [19].

In particular, manipulation participants' analyzer represents the manipulator who performs the manipulation, and whether the manipulator acts by him or has networks of other agents who helped him to execute the manipulation. Furthermore, the analyzer extracts the information describing benefits behind such manipulation, whether they accrue to the manipulator or to others. Finally, it checks whether the manipulator has any previous records or history of manipulations or violations. In total, the analyzer has 9 linguistic patterns to answer these questions and describe the manipulator and his social network profile.

The first three patterns show the agent who performed the violation and the manipulation activity type, as shown in figure 2. The patterns automatically analyze the sentence, extract the concept 'Robert J. Gallivan' and classify it under the <Person> sub-category. The concept 'defendant' is classified under the <LegalTitle> sub-category, the concepts 'breached a duty of trust and confidence' and 'insider trading activity' under the <Insider Trading> sub-category, and the concept 'the C&B consulting Firm' under 'organization'. Using the regular expression the analyzer retrieves the dates corresponding to the manipulation activity. The patterns automatically classify these sentences as the manipulator who performed the manipulation and map it to the 'ManipulationParticipants\Agent\AgentCharacteristics\Individual'. The line width and node sizes in a concept graph represent the global frequency counts of the extracted concepts from the document. For example, apparently the concepts 'Robert J. Gallivan' and 'breached a duty of trust and confidence' were mentioned in the document several times, represented by the thickness (Global count 5) of the line as shown in figure 2. In order to check whether the manipulator has a previous violation record, three patterns are developed to automatically analyze the complaint document and extracts the concepts that describe the manipulation history of the manipulator. The Commission found that Gallivan, who was affiliated with a broker-dealer at the time of the scheme, wilfully violated Section 17(a) of the Securities Act of 1933, in 1975, without admitting or denying the Commission's findings, Gallivan consented to the entry of a Commission order against him in the Proceeding File No. 3-4425.

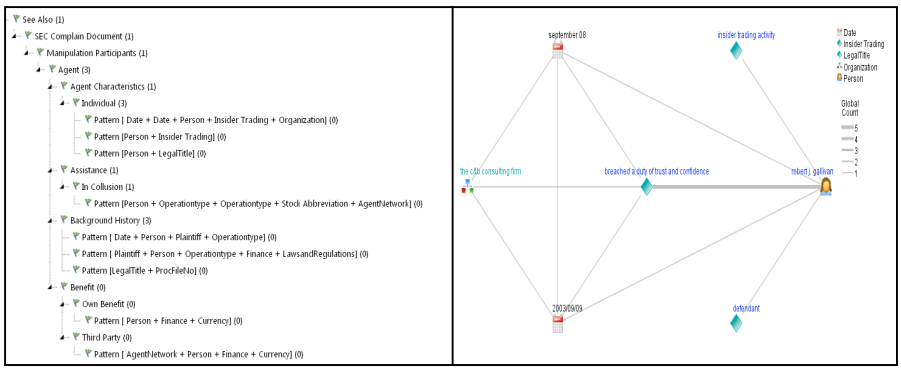


Fig 2 Manipulation Participant Analyzer

The last three patterns in the manipulation participants' analyzer are developed to extract the information that addresses the manipulator's social network, which helped him to violate the securities Harbour, Mid valley and Valencia securities (Target Assets). In this case Gallivan recommended the purchase of different stocks to his relatives, friend, and cousin to gain unlawful and combined profits reaching $58,453. The patterns automatically classified these sentences to the ontology class as the agent was in collusion with other agent networks (Assistance\In collusion). Furthermore, both the manipulator and his networks received benefits from the manipulation (Benefit\Own Benefit and Benefit\Third Party), as shown in figure 3.

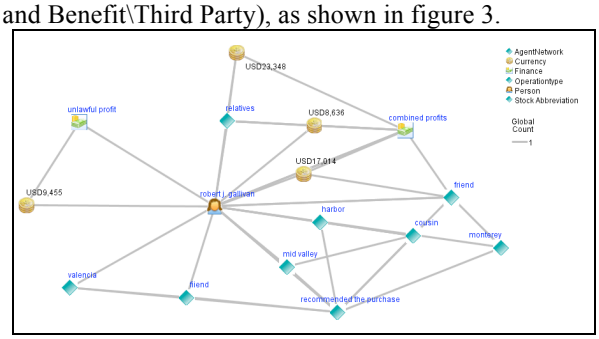


Fig 3 Agents' Social Network with Benefits

Timeline Manipulation Events and Actions Analyzer combine three analyzers 'Actions', 'Effects', and 'Time'. This analyser demonstrates the patterns developed for the three analyzers. The analysis indicates a strong relationship between the three analyzers because they describe the facts and nature of the manipulation activity executed by fraudsters. These actions could be related information-based activities such as obtaining non-public information and breach of confidence or trust, or could be trade-based activities such as buying and selling stocks to stimulate the market and violate prices. In this case, each action is associated with the temporal dimension and explains the period in which manipulator performed these manipulative activities.

These actions have an effect on the manipulated assets represented in direct or indirect benefits and unlawful profits. In this case, patterns and evidence such as legal, financial and economic unstructured information are used to trace the behavior of manipulators and show the consequences of their behavior on the market.

This analyzer contains 50 advanced linguistic patterns to extract information related to facts and the actions executed by the manipulator associated with the timeline. nine patterns are classified under 'Actions' classes , 20 patterns extract concepts related to the 'Effects' of manipulation , and 21 patterns are classified under the 'Time' class which describe the events before, during and after the fraud.

Figure 4 demonstrates the output of the 50 patterns developed for this analyzer applied to the 'insider trading' case study. The output shows the complexity of the manipulated activities executed by the agent. Most of the events are interconnected and interrelated, such as date, agreements and confidentiality, stocks prices, amount of purchases, manipulated assets, agent and his networks, amount of combined profits collected by the agent, the way of communication and meeting to obtain the non-

public material to take advantage, other evidence and patterns either legal or related to economic structure or trading used by the agent to violate the market.

In this case, the manipulator violated the prices of four stocks: Valencia Stock, Sun Country Stock, Mid valley Stock, and Harbor Stock. The manipulation actions between the four stocks are similar, as the manipulator 'Robert J. Gallivan' has breached a duty of trust and confidence of his company 'C&B Consulting Firm'. Based on his role in the company, the manipulator attended meetings and set up calls with companies and investors which gave him an opportunity to obtain non-public material. 'Robert J. Gallivan' used this information to buy these stocks and recommended them to his social network to combine profits from these trading transactions. Indeed, the agent agreed and signed that he would keep the matter confidential, but this was not the case and he breached this trust and violated the securities based on the insider information he acquired.

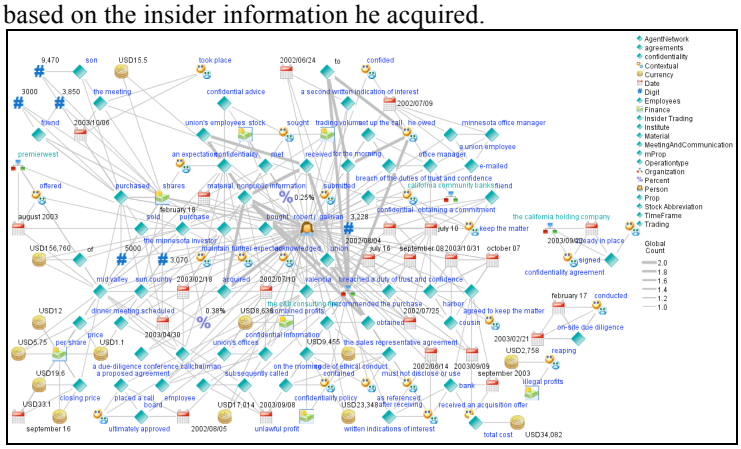


Fig 4 Timeline Manipulation Event and action Analyser Output

5 Conclusions

This paper contributes to market monitoring surveillance systems work. A linguistic based text mining approach is demonstrated for different market manipulation types based on the SEC litigation releases. The approach provides empirical evidence of how text mining could be integrated with the financial fraud ontology to improve the efficiency and effectiveness of extracting financial concepts. However, focusing on only a few case studies can potentially skew the outcome. This paper has limitations regarding the coverage of the cases and datasets.

In terms of future work, it is still possible to enhance and expand the cases to evaluate the text mining model by including new manipulation schemes and corresponding concepts on the basis of the SEC and other possible sources. For example, high frequency trading and stuff-quoting examples of possible stock-market manipulation cases should be included. Future work will pursue the full deployment of the text mining solution for the SEC litigation use as fraud knowledge management portal.

References

1. Aggarwal, R.K. and Guojun, W.U. Stock Market Manipulation -- Theory and Evidence. *Working Papers (Faculty) -- University of Michigan Business School*, 2003,
2. Diaz, D., Theodoulidis, B., and Sampaio, P. Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices. *Expert Systems with Applications 38*, 10 (2011), 12757–12771.
3. Diaz, D. and Theodoulidis, B. Financial Markets Monitoring and Surveillance: A Quote Stuffing Case Study. *Available at SSRN 2193636*, (2012).
4. Diaz, D., Zaki, M., Theodoulidis, B., and Sampaio, P. A Systematic Framework for the Analysis and Development of Financial Market Monitoring Systems. *2011 Annual SRII Global Conference*, Ieee (2011), 145–153.
5. Donoho, S. Early detection of insider trading in option markets. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.
6. Goldberg, H., Kirkland, J., Lee, D., Shyr, P., and Thakker, D. The NASD Securities Observation, News Analysis & Regulation System (SONAR). *Proceedings of the Fifteenth Conference on Innovative Applications of Artificial Intelligence*, (2003).
7. IBM. SPSS Modeler. 2013. http://www.ibm.com/software/analytics/spss/products/modeler/.
8. Kingston, J., Schafer, B., and Vandenberghe, W. Towards a financial fraud ontology: A legal modelling approach. *Artificial Intelligence and Law 12*, 4 (2004), 419–446.
9. Kirkland, J.D., Senator, T.E., and Hayden, J.J. Advanced-Detection System (ADS). *20*, 1 (1999), 55–68.
10. Laroque, R.B.; G. Using Previleged information to manipulate markets.pdf. *The quarterly Journal of Economics*, (1992).
11. Martin A. Rogoff. Legal Regulation of OTC market manipulation criticq proposal. *Finance 1*, (2012).
12. Mirchandani, V.K. and R. Increasing the ROI of Social Media Marketing REPRINT NUMBER. *MIT Slogan Management Review 54*, 1 (2012).
13. Mongkolnavin, J. and Tirapat, S. Marking the Close analysis in Thai Bond Market Surveillance using association rules. *Expert Systems with Applications 36*, 4 (2009), 8523–8527.
14. Polansky, S., Kulczak, M., and Fitzpatrick, L. NASD Market Surveillance Assessment and Recommendations. Final Report. *Achievement of Market Friendly Initiatives and Results Program (AMIR 2.0 Program)*, 2004. http://pdf.usaid.gov/pdf_docs/PNADB391.pdf.
15. Vila, J. Simple games of market manipulation. *Economics Letters 29*, (1989), 21–26.
16. Xia. Applying data mining in market abuse detection. 2007.
17. Zaki, M., Diaz, D., and Theodoulidis, B. Financial Market Service Architectures: A "Pump and Dump" Case Study. *2012 Annual SRII Global Conference*, (2012), 554–563.
18. Zaki, M., Theodoulidis, B., and Solís, D.D. "Stock-touting" through spam e-mails: a data mining case study. *Journal of Manufacturing Technology Management 22*, 6 (2011), 770–787.
19. Zaki, Mohamed and Theodoulidis, Babis. An Ontology for Monitoring and Surveillance in Financial Markets. *SSRN Electronic Journal*, (2013)

Predicting the impact of central bank communications on financial market investors' interest rate expectations

Andy Moniz[1] and Franciska de Jong[2,3]

[1] Erasmus Studio, Erasmus University, Rotterdam, The Netherlands
{moniz}@rsm.nl
[2] Erasmus Studio, Erasmus University, Rotterdam, The Netherlands
{fdejong}@eshcc.eur.nl
[3] Human Media Interaction, University of Twente, Enschede, The Netherlands
{f.m.g.dejong}@utwente.nl

Abstract. In this paper, we design an automated system that predicts the impact of central bank communications on investors' interest rate expectations. Our corpus is the Bank of England's '*Monetary Policy Committee Minutes'*. Prior studies suggest that effective communications can mitigate a financial crisis; ineffective communications may exacerbate one. The system described here works in four phases. First, the system detects salient aspects associated with economic growth, prices, interest rates and bank lending using information from Wikipedia. These *economic aspects* are detected using the *TextRank* link analysis algorithm. A multinomial Naive Bayesian model then classifies document sentences to these aspects. The second phase measures sentiment using a count of terms from the *General Inquirer* dictionary. The third phase employs Latent Dirichlet Allocation (LDA) to infer topic clusters that may acts as intensifiers/diminishers for the economic aspects. Finally, an ensemble tree combines the phases to predict the impact of the communications on financial market interest rates.

Keywords: sentiment analysis·text mining·link analysis·financial markets

1 Introduction

Post the global financial crisis, there has been a dramatic change in the use of central bank communications as a central bank policy instrument [1,2]. Central banks communicate qualitative information to the financial market through statements, minutes, speeches, and published reports [3]. Communication is an important tool that a central bank can use to avert a crisis, by providing investors with its assessment of the risks and the measures it views as necessary to reduce those risks within the economy [1], [4]. Previous studies suggest that effective central bank communications can mitigate and potentially prevent a financial crisis; ineffective communications may exacerbate one [1], [5]. In [4], the Swedish central bank, the Riksbank, is criticized because its communications were "not clear or strong enough" leading up to the global financial crisis, such that the bank's information went "unnoticed" [1]. In this paper, we design

an automated system that predicts the impact of central bank communications on interest rate expectations, as derived via financial market patterns. For the purposes of this study, we analyze economic sentiment, as expressed in the *'Monetary Policy Committee Minutes'* [2] published by the Bank of England, that details its monthly interest rate decisions.

Financial markets scrutinize central bank communications for "*clues and shades of meaning about its assessment of the economy and the direction of where economic policy may be heading"* [1]. As a prediction task, the measurement and evaluation of sentiment is challenging due to the complexities and subtleties of interpreting bank communications [1]. The formation of economic policy is a balancing act between achieving high economic growth and financial stability, while targeting low inflation [2]. The relative importance of these objectives is dynamic, and varies depending on the prevailing economic conditions [2]. For example under benign economic conditions, high inflation may be construed by financial market investors as a negative signal for the direction of future interest rates. During the financial crisis of 2007-2009, high inflation was considered to be a positive signal by effectively lowering interest rates[1] [6]. This motivates a need for fine-grained sentiment analysis, to automatically detect economic aspects and predict the central bank sentiment expressed towards these aspects [7]. Such a model would provide investors with an automated system to decipher the complexities and interactions of *economic aspects*, to interpret the consequences of these interactions for the future path of interest rates, and to incorporate the information into their investment decisions. For a central bank, such a model would provide it with the ability to predict the impact of its economic policies on the financial markets. The resulting 'price discovery' process [2] may promote a more efficient functioning of financial markets.

Our approach consists of four phases. First, the system detects salient references to economic aspects associated with economic growth, prices, interest rates and bank lending and employs a multinomial Naive Bayesian model to classify sentences within documents. Economic aspects are identified in a pre-processing step, that employs a link analysis using the TextRank algorithm [8,9]. The second phase measures sentiment expressed for the economic aspects, using a count of terms from the *General Inquirer* dictionary [17]. The third phase employs Latent Dirichlet Allocation (LDA) to infer intensifiers/diminishers that may change the meaning of the economic aspects and economic sentiment [10],[7]. Specifically, the model categorizes whether the magnitude of the economic aspects has 'intensified' or 'diminished' over time [11,12]. We refer to the resulting topic clusters as *directional topic clusters*. Finally, an ensemble tree combines the model components to predict the impact of the communications on financial market interest rates over the following day.

[1] The real interest rate is the rate of interest a borrower expects to pay on debt after allowing for inflation and is equal to the nominal interest rate (set by the central bank) minus the rate of inflation [2]

The rest of this paper is structured as follows. Section 2 draws on literature from the fields of economics and discusses the implications for sentiment analysis and keyword detection. Section 3 models the individual components of the system. Section 4 outlines the corpus of central bank communications, provides an evaluation of the model components and then discusses the results. Section 5 concludes and suggests avenues for future research.

2 Related Work

2.1 Background: central bank research

Post the financial crisis, several central banks have identified communications, particularly 'enhanced forward guidance', as an important policy instrument within their economic toolkit [1,2]. Effective communications enhance a central bank's public transparency, accountability and credibility [13], which in turn aids its ability to implement economic policies [14]. To date, there has been little research into text mining of central bank communications. In [14], the impact of different types of communications (press releases, speeches, interviews, and news conferences) are analyzed to determine which media sources impact interest rate expectations. The analysis does not, however, classify the language used in the documents. In [3], a term counting approach is adopted to analyze the sentiment contained within the meeting minutes of the US central bank (the Federal Reserve). In [3] and [15] Latent Semantic Analysis is employed to analyze the sentiment contained within the Bank of Canada's minutes. The intention of this study is to design a fine-grained sentiment analysis approach to analyze the impact of central bank communications on financial market investors. To our knowledge, this remains an unexplored avenue of research.

2.2 Background: sentiment analysis

Traditionally, fine-grained sentiment analysis has been researched for the classification of online user reviews of products and movies [16]. Readers are often not only interested in the general sentiment towards an aspect but also a detailed opinion analysis for each of these aspects [7]. Evaluation is conducted by comparing model classifications versus ratings provided by users. The evaluation of economic sentiment is arguably a harder task, due to the lack of a clearly defined outcome to assess model performance. For example, which economic variable should a model's predictions be evaluated against? The relative importance of the aspects (e.g. economic growth/ inflation/interest rates) is subjective, may vary over time, and the measurement of the aspects is only known with significant time delay.

The traditional approach to text-mining within the field of finance is to count terms using the General Inquirer dictionary [17,18]. The dictionary classifies words according to multiple categories, including 1,915 positive words and 2,291 negative words. The *General Inquirer* was developed for psychology and sociology research and

while it is used for text mining within the field of finance, little research has been conducted as to its suitability within finance [19]. Aspects that are frequently mentioned in central bank communications, such as the terms 'employment', 'unemployment' and 'growth', are not classified by the *General Inquirer* dictionary. Adjectives are often needed before investors can interpret the patterns in the economy to form their interest rate expectations [3]. Furthermore, the terms 'inflation' and 'low' are classified as negative by the dictionary, yet 'low inflation' is a positive characteristic and indeed achieving this is a central bank's core objective [2]. The terms 'fall' and 'decline' are classified as negative terms in the *General Inquirer* dictionary, yet the opposite terms 'rise' and 'increase' are not classified at all.

2.3 Background: keyword detection

Graph-based algorithms have received much attention [8] as an approach to keyphrase extraction and are considered to be state-of-the-art unsupervised methods [20]. In a graph representation of a document, nodes are words or phrases, and edges represent co-occurrence or semantic relations. The underlying assumption is that all words in the text have some relationship to all other words in the text. Such an approach is statistical, because it links all co-occurring terms without considering their meaning or function in text. Centrality is often used to estimate the importance of a word in a document [22], and is a way of deciding on the importance of a vertex within a graph that takes into account global information recursively computed from the entire graph, rather than relying only on local vertex-specific information [23]. The main advantage of such a representation is that selected terms are independent of their language [21].

3 Model to predict changes in investors' expectations

In this section we describe the four phases of the system. First, the system detects salient references to economic aspects and employs a multinomial Naive Bayesian model to classify sentences within documents. The second phase measures sentiment expressed for the economic aspects, using a count of terms from the *General Inquirer* dictionary. The third phase employs a LDA model and categorizes whether the magnitude of the economic aspects has 'intensified' or 'diminished' [11,12]. Finally, an ensemble tree combines the model components to predict the impact of the communications on financial market interest rates over the following day.

3.1 Aspect detection

In [3] it is shown that *tf-idf* weighting selects infrequent terms that relate to major news events or economic shocks. Our approach is intended to detect the common economic themes that are discussed in central bank communications, that are more likely to influence investors' interest rate expectations on a day-to-day basis [2]. To determine salient references, we employ a link analysis approach that detects the most frequently mentioned terms within two Wikipedia pages on Central Banking and In-

flation. The model employs TextRank [8], a ranking algorithm based on the concept of eigenvector centrality, to compute the importance of the nodes in the graph. Each vertex corresponds to a word. A weight, w_{ij}, is assigned to the edge connecting the two vertices, v_i and v_j. The goal is to compute the score of each vertex, which reflects its importance, and use the word types that correspond to the highest scored vertices to form keywords for the text [23]. The score for v_i, $S(v_i)$, is initialized with a default value and is computed in an iterative manner until convergence using recursive formula shown in Equation (1).

$$S(v_i) = (1-d) + d \times \sum_{v_j \epsilon Adj(v_i)} \frac{w_{ji}}{\sum_{v_k \epsilon Adj(v_j)} w_{jk}} S(v_j) \tag{1}$$

where $Adj(v_i)$ denotes v_i's neighbors and d is the damping factor set to 0.85 [8]. Figure 1 displays the resulting clustering of terms. The size of each node is directly proportional to the TextRank score of the respective economic aspect.

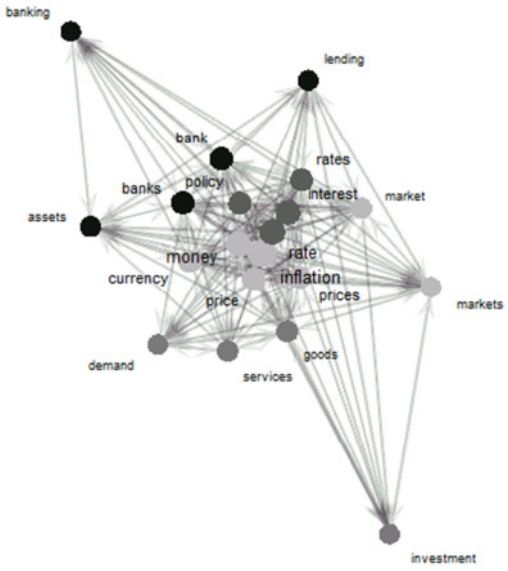


Fig. 1. Link analysis of frequently occurring terms. Different nodes colors reflect different communities identified using the Clauset-Newman-Moore algorithm.

We define economic aspects by employing a greedy algorithm to detect communities of terms within the network [31]. The *economic growth aspect* detects the frequency of the terms: *'demand'*, *'goods'*, *'services'*, *'investment'*. The *prices aspect* detects the terms: *'inflation'*, *'prices'* , *'money'*, *'markets'*, *'currency'*. The *interest rate aspect* detects the occurrence of: *'interest'*, *'rates'*, *'policy'* and a *bank lending aspect* detects the terms: *'banks'*, *'lending'* and *'assets'*. It is not surprising to see these terms appear in the link analysis, given a central bank's remit is to maintain price and financial stability. The choice of terms is consistent with the text mining

research of [3] which identifies 'growth', 'price', 'rate', and 'econom' as the most frequently occurring terms for the US economy. Using the four economic aspects, the system employs a multinomial Naive Bayesian model [24] to categorize sentences within each document. The resulting classification labels form the basis upon which sentiment analysis is applied.

3.2 Polarity detection

In the second phase, the model computes a measure of economic sentiment associated with each of the economic aspects. We measure polarity by counting the number of positive (P) versus negative (N) terms, (P − N)/(P + N) identified using the *General Inquirer* dictionary [17]. In line with [16], our goal is not to show that a term counting method can perform as well as a Machine Learning method, but to provide a baseline methodology to measure central bank sentiment and to draw attention to the limitations of the approach that is widely adopted by text mining studies in the field of finance as indicated in Section 2.2. The sentiment metrics that are associated with the economic aspects: economic growth, prices, interest rate and bank lending are labelled $Tone_{growth}$, $Tone_{prices}$, $Tone_{interest\_rates}$ and $Tone_{bank\_lending}$ respectively. A fifth sentiment metric, $Tone_{overall}$, is computed to measure the polarity associated with the overall document, without conditioning upon the *economic aspects*. The five sentiment metrics are included as separate components within the ensemble tree.

3.3 Detection of LDA directional topic clusters

Next we extend the baseline term-counting method by taking intensifiers and diminishers into account [11,12]. These are terms that change the degree of the expressed sentiment in a document (see Section 2.2). In the case of central bank communications, the terms describe how economic aspects have changed over time. We employ an implementation of LDA [10], and represent each document as a probability distribution over latent topics, where each topic is modeled by a probability distribution of words. In [7], LDA is found to capture the global topics in documents, to the extent that topics do not represent ratable aspects associated with individual documents, but define clusterings of the documents into specific types. For the purposes of training the LDA model, we consider each sentence within each central bank communication to be a separate document. This increases the sample size of the dataset (see Section 4.1) and is intended to improve the robustness of the LDA model for statistical inference. We implement standard settings for LDA hyper-parameters, $\alpha = 50/K$ and $\beta=.01$, where the number of topics K is set to 20 [25]. We manually annotate two of the topic clusters that capture 'directional' information [1] and appear to act as intensifiers/diminishers of meaning. We label the clusters *directional topic clusters*. Table 1 identifies the top terms associated with the two clusters. Representative words are the highest probability document terms for each topic cluster.

Table 1. Representative document terms associated with the directional topic clusters

'intensifier cluster'		*'diminisher cluster'*	
word	**prob.**	**word**	**prob.**
increase	0.150	moderated	0.190
strong	0.107	lower	0.161
accelerate	0.081	downwards	0.123
strength	0.063	difficult	0.102
support	0.058	less	0.070

Next for each central bank communication the LDA model infers the probabilities associated with the 'intensifier' and 'diminisher' clusters within each of the three economic aspects detected by the Naïve Bayesian classifier. The output of the model is a vector of six topic probabilities that proxy the central bank's assessment that the economic aspects are intensifying/diminishing. We label the model *directional LDA model* and the respective probability vectors: $\text{Topic}_{\text{growth}_{\uparrow}}$, $\text{Topic}_{\text{prices}_{\uparrow}}$, $\text{Topic}_{\text{interest\_rates}_{\uparrow}}$ and $\text{Topic}_{\text{bank\_lending}_{\uparrow}}$ if the economic aspects are increasing and $\text{Topic}_{\text{growth}_{\downarrow}}$, $\text{Topic}_{\text{prices}_{\downarrow}}$, $\text{Topic}_{\text{interest\_rates}_{\downarrow}}$ and $\text{Topic}_{\text{bank\_lending}_{\downarrow}}$ if the economic aspects are decreasing. We include the topic probabilities as components within the ensemble tree.

4 Experiments

In this section we discuss the corpus of central bank communications and describe the investor patterns data used to evaluate the impact of the central bank communications on investors' interest rate expectations. We then outline the evaluation of the ensemble classification tree, present the results and provide a discussion.

4.1 Data

We choose to analyze the interest rate minutes of the *Bank of England*. As cited in [3], central bank minutes are closely watched by investors to gauge the future direction of economic policies. Similar datasets for the US and Canadian central banks' minutes are examined in [3] and [15]. The Bank of England announces the level of UK interest rates on the first Thursday of every month. The details that underpin this decision are only provided two weeks later and are published in the Bank of England's *'Monetary Policy Committee Minutes'*. The communications are interesting to analyze, because changes in investors' expectations on the day of the central bank communication may be attributed to the qualitative information contained within the meeting minutes rather than the interest rate decision announced two weeks before. Minutes typically include summaries of committee members' views on economic conditions and discuss the rationale for their interest rate decisions [26]. The central bank's minutes are, on average, 12 pages long (including a header page), and contain around 55 bullet points, typically with 5 sentences in each bullet. The documents are available from 1997, the year when Parliament voted to give the Bank of England operational independence from the UK government. We retrieve all meeting minutes

available between July 1997-March 2014[2] to create a corpus that consists of 199 documents. For the purposes of aspect detection and to train the LDA model, we remove the header page and define a document as an individual sentence within each of the meeting minutes. This expands the corpus to a collection of 53,195 documents.

To evaluate the ensemble tree's predictions we utilize information obtained from financial market patterns. Interest rate futures contracts are financial instruments that enable investors to insure against or speculate on uncertainty about the future level of interest rates [27]. Changes in the price of the futures contracts therefore reflect changes in investors' views on the future direction in central bank interest rates. Investors' interest rate expectations for the following three, six and twelve months are derived and published daily by the Bank of England. We utilize investors' twelve month ahead forecasts. This data series has the greatest data coverage compared to the three and six month series. Furthermore, the twelve month forecast horizon is consistent with the time horizon over which that the Bank of England conducts its economic policies [2]. To isolate the effect of the central bank communication on investors' expectations, we compute the percentage change in the interest rate futures contract, as measured from the close of business on the day of the communication announcement until the close of business one day after. This narrow time window helps to minimize the influence on investors' interest rate expectations from other financial market factors that may occur at the same time [28].

4.2 Experiment setup

We design the evaluation of the prediction model in stages to enhance our understanding of the model components. For a baseline, we evaluate the system's predictions by using only the tone of the overall document (see Section 3.2). The approach does not take into account individual economic aspects or diminishers/intensifiers [11,12]. We label the model *naïve tone*. This approach is consistent with the methodology adopted by the financial literature [18]. Next we compare the outcomes of an ensemble model that combines the tone associated with each of the economic aspects: economic growth, inflation and interest rates (see Section 3.2). We label this the *economic aspects model*. A third model compares the outcomes from an ensemble model that combines the tone of eight directional economic aspects, that combines the intensifiers/diminishers associated with the four economic aspects (see Section 3.3). We label this the *directional LDA model*. Finally, we combine the components in a single ensemble tree and refer to the system as the *joint aspect-polarity* model.

Learning and prediction is performed using an ensemble tree. The goal of ensemble methods is to combine the predictions of several models built with a given learning algorithm in order to improve generalizability and robustness over a single model. We use the Random Forest algorithm [30], that employs a diverse set of classifiers by introducing randomness into the classifier construction. Experiments were validated

[2] Central bank communications announced in August 1997 were excluded from the analysis because the communication document was not readily available in a machine readable format.

using five-fold cross validation in which the dataset is broken into five equal sized sets; the classifier is trained on four datasets and tested on the remaining dataset. The process is repeated five times and we calculate the average across folds. For evaluation, we select Mean Absolute Error (MAE), Root Mean Squared Error and Spearman's rho (ρ). We also examine Spearman's rho since prediction may be considered to be a ranking task. The formulae are displayed in Equation (2) below.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |O_i - E_i| \quad , \quad RMSE = \left[\frac{1}{n}\sum_{i=1}^{n} |O_i - E_i|\right]^2 \quad , \quad \rho = 1 - \frac{6\sum(O_i - E_i)^2}{n(n^2-1)} \tag{2}$$

where E_i is the model's predicted value, O_i is the realized value, and n is the number of observations. MAE measures the average magnitude of the forecast errors without considering direction; RMSE penalizes errors and gives a relatively high weight to large errors. A smaller value of MAE or RMSE indicates a more accurate prediction. Spearman's rho is a non-parametric measure of the degree of linear association between the predicted and realized values, and is bound between the range -1 to +1 [29]. A positive Spearman's rho indicates the model's predictive ability; a negative value indicates a poor model fit.

4.3 Experiment results

The evaluation metrics from the model components are shown in Table 2.

Table 2. Evaluation of the model components

Model	MAE	RMSE	ρ
naive tone	0.022	0.016	-0.187 ***
economic aspects	0.018	0.013	-0.044
directional LDA	0.019	0.014	0.041
joint polarity model	0.015	0.011	0.034

The asterisks provide the level of significance where *** indicates that the model's predictions versus forecasts are statistically, negatively significant at the 0.1% level.

The naïve tone model, which is similar to the approach commonly adopted by text mining studies in the field of finance, shows the worst performance. It exhibits the highest MAE and RMSE. The rank correlation of the model's forecasts with realized changes in investors' interest rate expectations is negative and is highly statistically negative, implying that documents that are predicted to have a positive/negative impact on investors' interest rate expectations end up having the reverse effect. The economic aspects and directional LDA models exhibit monotonic decreases in MAE and RMSE, suggesting a slight improvement in the model fit. Spearman's rho, however, is again negative, albeit to a lesser extent. Finally, the joint aspect-polarity model, that includes all model components in the ensemble tree, displays the lowest MAE and RMSE. The mildly positive Spearman's rho is consistent with previous studies within the field of finance. As cited in [19], many factors influence the financial mar-

kets; a low, positive correlation provides sufficient comfort of the model's predictive power.

4.4 Discussion

One interpretation of the experiment results is that multiple aspects are needed to improve the accuracy of the system. The positive Spearman's rho for the joint model versus the negative Spearman's rho for the naïve tone, economic aspects and directional LDA models may be indicative of a non-linear relationship between the components that is only evident when the models are combined rather than considered in isolation. One of the strengths of a regression tree is that it does not assume a functional form, allowing it to detect interactions between model components. To aid our understanding of prediction in the joint model, Figure 2 displays the decision tree results for one of the folds. The values in the grey boxes provide the predicted percentage change in investors' interest rate expectations associated with the sentiment contained within the central bank communication. A positive value indicates that the impact is expected to lead to an increase in investors' interest rate expectations, while a negative value indicates an expected decrease in interest rate expectations.

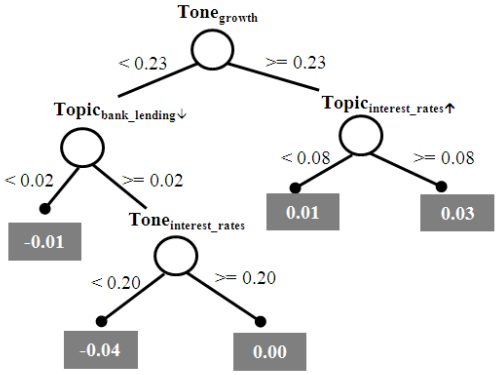


Fig. 2. Example decision tree from one of the folds

The regression tree identifies the interaction between the *directional topic clusters* and *Tone* measures. The primary decision in the decision tree is central bank sentiment towards economic growth. The right hand path indicates that if a central bank communication emphasizes positive economic growth and discusses interest rate increases, the model predicts that investors' expectations of future interest rates will rise by 3%. The left hand path predicts that if a central bank tone towards economic growth is low, declining bank lending and the tone towards interest rates is negative, investors will reduce their expectations of future interest rates by 4%.

5 Conclusion

The goal of central bank communication is to make messages as clear, simple and understandable as possible to a wide range of audiences [1]. In this study, we focus of one specific audience, namely financial market investors. Investors play a key role for the implementation of a central bank's economic policies [1,2]. The outcome of our study may feed the design of a system that can predict the impact of central bank communication on formation of investors' interest rate expectations. The results of the joint aspect-polarity model suggest that investors may benefit by incorporating a measure of central bank sentiment to forecast interest rates.

In this study we evaluate model performance using prices from financial market instruments. The market price of an interest rate contract implicitly measures the average investor's interest rate expectations [27]. It is also possible to compute an 'implied probability distribution' of those expectations [27]. In future work we plan to evaluate a range of metrics, including the dispersion of the expectations as a proxy of investor uncertainty. Post the 2007–09 financial crisis, central banks have broadened the range of their communication, including the use of social media, live broadcasts, podcasts and blogs, to deliver their messages quickly and efficiently [1]. In future research, a wider range of central bank communications, including those expressed via social media, will be integrated into our study. We also intend to examine alternative approaches to select economic aspects, including dynamic approaches that reflect the usage of terms as central bank communications change over time.

Acknowledgement

The research leading to these results has partially been supported by the Dutch national program COMMIT.

References

1. Vayid, I. Central Bank Communications Before, During and After the Crisis: From Open-Market Operations to Open-Mouth Policy, Bank of Canada Working Paper (2013)
2. Bank of England. Monetary policy trade-offs and forward guidance (2013)
3. Boukus, E., Rosenberg, J., V. The information content of FOMC minutes (2006)
4. Meyersson, P., Karlberg, P.P, A Journey in Communication: the Case of the Sveriges Riksbank SNS Förlag (2012)
5. Viñals, J. Lessons from the Crisis for Central Banks. IMF Speech (2010)
6. Danthine, J, P. Causes and consequences of low interest rates. Speech by Mr Jean-Pierre Danthine, Vice Chairman of the Governing Board of the Swiss National Bank, at the Swisscanto Market Outlook 2014, Lausanne (2013)
7. Titov, I, McDonald, R., T. Modeling online reviews with multi-grain topic models. Proceeding of the 17th WWW (2008)
8. Mihalcea, R., and Tarau, P. TextRank: Bringing order into texts. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (2004)

9. Brin, S., Page, L. The anatomy of a large-scale hypertextual web search engine. Comput. Networks ISDN Systems, 33:107–117 (1998)
10. Blei, D., M., Ng, A., Jordan, M., I. Latent Dirichlet Allocation, Journal of Machine Learning Research 3, 993-1022 (2003)
11. Kennedy, A., Inkpen, D. Sentiment Classification of Movie Reviews using Contextual Valence Shifters. Computational Intelligence, vol.22(2), pp.110-125 (2006)
12. Polanyi, L., Zaenen, A. Contextual Valence Shifters, AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (2004)
13. Carney, M. Panel discussion comments to the BIS Conference on The Future of Central Banking under Post-Crisis Mandates (2010)
14. Fay, C., Gravelle, T. "Has the Inclusion of Forward-Looking Statements in Monetary Policy Communications Made the Bank of Canada More Transparent?" Bank of Canada Discussion Paper (2010)
15. Hendry, S., Madeley, A. Text Mining and the Information Content of Bank of Canada Communications, Bank of Canada Working Paper (2010)
16. Pang, B., Lee, L., Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of EMNLP-02 (2002)
17. Stone, P., Dumphy, D. C., Smith, M. S., and Ogilvie, D. M. The General Inquirer: A Computer Approach to Content Analysis. The MIT Press. (1966)
18. Tetlock, P., Saar-Tsechansky, M. Macskassy, S. More Than Words: Quantifying Language to Measure Firms' Fundamentals. Journal of Finance, Vol. LXIII, No. 3, 1437-1467 (2008)
19. Loughran, T., McDonald, B. When is a liability not a liability? Textual analysis, dictionaries and 10Ks. Journal of Finance 66, 35–65 (2010)
20. Liu, F. , Pennell, D., Liu, F., Liu, Y. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American (2009)
21. Litvak, M., Last, M. Graph-Based Keyword Extraction for Single-Document Summarization. Proceedings of the 2nd Workshop on Multi-source, Multilingual Information Extraction and Summarization, Coling 2008, pp. 17-24. Association for Computational Linguistics (2008)
22. Opsahl, T., Agneessens, F., Skvoretz, J. Node centrality in weighted networks: Generalizing degree and shortest paths (2010)
23. Boudin, F. A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction (2013)
24. McCallum, A., and Nigam, K. A comparison of event models for naive Bayes text classification. AAAI-98 Workshop on Learning for Text Categorization (1998)
25. Griffiths, T. L., Steyvers, M. Finding scientific topics. Proceedings of the National Academy of Science, 101, 5228-5235 (2004)
26. Danker, D., J., and Luecke, M., M. Background on FOMC Meeting Minutes, Federal Reserve Bulletin 175-179 (2005)
27. Clews, R., N. Panigirtzoglou, Proudman, J. Recent developments in extracting information from options markets, Bank of England Quarterly Bulletin (2000)
28. Mackinlay, A., C. Event Studies in Economics and Finance. Journal of Economic Literature (1997)
29. Maritz, J., S. Distribution-free statistical methods. Chapman and Hall, (1984)
30. Breiman, L. Random Forests. Machine Learning (2001)
31. Clauset, A., Newman, M. E. J., and Moore, C. Finding community structure in very large networks. Physical Review (2004)

Predicting stocks returns correlations based on unstructured data sources

Mateusz Radzimski, José Luis Sánchez-Cervantes, José Luis López Cuadrado, Ángel García-Crespo

Departamento de Informática
Universidad Carlos III de Madrid, Spain
{mradzims, joseluis.sanchez, joseluis.lopez.cuadrado, angel.garcia}@uc3m.es

Abstract. The recent outbreak of information demand for financial investment management forces to look for novel ways of quantitative data analysis. Relying only on traditional data sources means to loose the edge over the competence and become irrelevant in the future. On the other hand, the Big Data and Data Analytics trends are getting traction in financial domain and are being sought as highly beneficial in the long term. This paper presents an approach for forecasting stock correlations based on big volumes of unstructured and noisy data. We evaluate the prediction model and demonstrate its viability for certain industrial sectors.

Keywords: quantitative analysis, unstructured data analysis, financial decision-making, data science.

1 Introduction

The recent dynamics of the data created on the Web can only be described as a massive data deluge. The expansion of the "digital universe" follows the predictions [1] and it doubles almost every two years [2]. The major part of this data is unstructured and has a form of videos, photos, news or other social media content. Such data can provide valuable insights, but are much more difficult to analyze and interpret correctly. The recent outbreak of Big Data technologies and novel approaches to data extraction offer new ways of understanding and dealing with such huge volumes of unstructured data. Combining that with structured information sources, such as financial statements or price evolution time series can improve decision making process in the financial domain. Applying data mining to unstructured financial data can reveal hidden correlations or even predict relevant economic indicators.

In this work we analyze the unstructured data in a form of financial news and we measure the impact of the news on the periodic returns of the stocks of analyzed companies. By examining the co-occurrence of the companies we estimate the correlation of the daily returns time series between company pairs. The result is analyzed in the context of portfolio management in order to improve the diversification of stocks and better spread the risk of financial investments by avoiding companies with high correlation of stock prices evolution.

2 Related Work

There are several initiatives related with predicting financial results based on unstructured data sources. Some of these works have obtained outstanding results through applying semantic technologies. In this section some of this works are described briefly.

An experimental platform for the evaluation of various approaches of automatic sentiment analysis in financial texts was presented in [3]. The platform provides an experimental environment, which enables the user to quickly assess the behavior of various algorithms for sentiment detection in given text from an arbitrary HTML source. There is a basic sentiment word tagger and country tagger available to facilitate the preview of text under analysis. Development of the platform and the algorithms it hosts is still in progress and is intended for experimental purposes only. However, as it is implemented as a publicly available Web application, it can find its use also in general public with an interest in sentiment revealing parts of HTML sources. Finally, authors plan to tackle on their platform, the sarcasm detection. They expect to experiment with a number of natural language analysis and machine learning tools for this purpose and with combinations of both.

In [4], the positive sentiment probability as a new indicator to be used in predictive sentiment analysis in finance was proposed. By using the Granger causality test they show that sentiment polarity (positive and negative sentiment) can indicate stock price movements a few days in advance. The authors adapted the Support Vector Machine classification mechanism to categorize tweets into three sentiment categories (positive, negative and neutral), resulting in improved predictive power of the classifier in the stock market application. Their study indicates that changes in the values of positive sentiment probability can predict a similar movement in the stock closing price in situations where stock closing prices have many variations or a big fall.

In [5] a knowledge-based approach for extracting investor sentiment directly from web sources was presented. This approach performs a semantic analysis that starts on the word and sentence level. The authors employ ontology-guided and rule-based Web information extraction based on domain expertise and linguistic knowledge. Furthermore, they evaluate their approach against standard machine learning approaches. A portfolio selection test using extracted sentiments provides evidence for the economic utility of investor sentiments from web blogs.

A service oriented stream mining workflow for sentiment classification through active learning was presented in [6]. In the context of this use case, authors present the general idea of active learning (AL) as well as an empirical evaluation of several active learning methods on a stream of opinionated Twitter posts. The preliminary experiments showed that AL helps significantly when only a few tweets (e.g., 100–200) are labeled. After 200 tweets are labeled, the accuracy of the SVM-AL-Clust algorithm is 7.5% higher when compared to the random selection policy.

Unfortunately, when more and more tweets are labeled, the differences between the evaluated algorithms (including random) diminish.

The Web of data promotes the idea that more and more data are interconnected. A step towards this goal is to bring more structured annotations to existing documents using common vocabularies or ontologies. Unstructured and semi-structured texts such as scientific, medical or news articles as well as forum and archived mailing list threads or (micro-) blog posts can hence be semantically annotated. In this sense there are various initiatives for extracting and analyzing information in unstructured, semi-structured or structured graphs forms. Some of these initiatives are described bellow.

In [7], an experimental evaluation of human driven named entity extraction performed by the Named Entity Recognition and Disambiguation (NERD) Web application was presented. Their evaluation was performed considering precision of Named Entities extraction, precision of the classification of the information unit into categories, precision of the disambiguation of the Named Entity with Web resources and the relevant score. Experiment results showed the strengths and weaknesses of five different tools. Furthermore, AlchemyAPI seems the best solution to extract named entities and to categorize them in a deep ontology. Through the ability to infer data from the LOD cloud, DBpedia Spotlight and Zemanta, they are able to assign meaningful URIs to the extracted concepts. Finally, experiments are polarized using the authority as a key selection in the data choice and grouped in similar categories.

In [8], an infrastructure that converts continuously acquired HTML documents into a stream of plain text documents was presented. The work presented by authors consists of RSS readers for data acquisition from different Web sites, a duplicate removal component, and a novel content extraction algorithm, which is efficient, unsupervised, and language-independent. The core of the proposed content extraction algorithm is a simple data structure called URL Tree. The performance of the algorithm was evaluated in a stream setting on a time-stamped semi-automatically annotated dataset, which was made publicly available. They compared the performance of URL Tree with that of several open source content extraction algorithms. The evaluation results show that our stream-based algorithm already starts outperforming the other algorithms after only 10 to 100 documents from a specific domain.

The analysis of large graphs plays a prominent role in various fields of research and is relevant in many important application areas. For this reason, work [9] presents state-of-the-art report that examines the survey of available techniques for the visual analysis of large graphs. In their work, authors discuss various graph algorithmic aspects useful for the different stages of the visual graph analysis process. They also present main open research challenges in this field.

An intercompany network in which social network analysis techniques are employed in order to identify a set of attributes from the network structure was presented in [10]. The network attributes are used in a machine learning process to predict the company revenue relation (CRR) that is based on two companies' relative quantitative financial data. The origin of research lies in exploiting the large volumes of online business news, as they provide an opportunity to explore various aspects of companies. In particular, work [10] is similar to our initiative, however we present important differences which are: (i) different methodology towards the data sources:

our data is based on the news from the web, while they use sources, which are already annotated (Yahoo Finance). For instance, the authors already know that a news article corresponds to the company X and then mention companies Y & Z. Based on that they create a directed graph connecting companies as follows X -> Y & X -> Z. In our approach we rely on the simple co-occurrence of companies and stocks (based on the named entity recognition process) without knowing exactly to what company the news article belongs. (ii) We add the temporal aspect to the data, while they analyze the whole 3 quarters together, in an aggregated manner. (iii) We analyze the correlation between the daily returns, while they calculate the company revenue relation.

These initiatives offer alternatives of solution to different financial situations as: sentiment analysis from twitter, text mining from different resources such as blogs, news, i.e. unstructured and semi-structured information. In comparative with the previously mentioned proposals, the main idea of our approach is based in obtaining large sets of unstructured financial information from financial news with the aim of extracting relevant knowledge and predicting financial correlations.

3 Extracting Relations from Unstructured Data

The unstructured data used in this work is based on the analysis of over 200 news sources, coming mostly from financial news services and financial blogs, but also from general news broadcasters, such as BBC News or CNN. The data was collected from the period of January 2013 – December 2013. The entire dataset was created within the project FIRST[1]. The creation of the dataset consisted of various steps pipelined together, such as: news acquisition, duplicate filtering, boilerplate removal and named entity recognition. The whole process has been described in [3], [11] and [12]. Although the original FIRST project aimed at sentiment extraction from financial news and performed further information extraction steps, for this work we use only intermediate result, which consist of occurrence of companies and stocks in the financial news.

The preprocessed financial news dataset contains around 2,300,000 distinct news articles, containing over 24,800,000 annotations of 6,000 named entities. For the sake of the further experiments, we focused on companies from the S&P500 index, consisting of the 500 biggest and most liquid stocks from the US market. The initial list of 500 companies has been reduced to 395 companies, after filtering companies with insignificantly low or no coverage in the whole dataset.

This dataset is a starting point in creating a network of relationships between companies. Assuming that there is a relation of some kind when two companies appear in the same news article, we started from creating a graph representation of all co-occurrences of companies in the news. Capturing such relation should also consider temporal aspect of the news data. This is due to the fact that each article describes certain aspects that are relevant in the moment of publishing and its impact fades with time. Therefore the model takes into account the temporal aspect as well.

[1] http://project-first.eu

From the preprocessed financial news data we extracted a list of named entities appearing in each news article. Based on that we created a graph of relationships $G = \{V, E\}$, where V is a set of vertices each representing different company and E is a set of edges, each representing co-occurrence of two companies when they both appear in a single news article. As a result we obtained an undirected, weighted, temporal graph, where weights $w(e)$ represent how many times two companies co-occur in news, as represented by edge e, and temporal function $t(e)$ associates date d to every edge of the graph. Figure 1 presents an example graph based on three articles A_1, A_2 and A_3 mentioning the following companies: A_1: $\{C_1, C_2, C_3\}$, A_2: $\{C_2, C_3, C_4\}$, A_3: $\{C_3, C_4\}$.

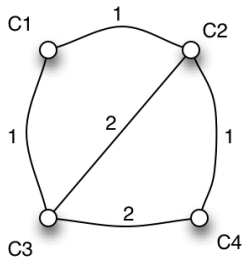


Figure 1: Example of companies' relationship graph with edge weights

The co-occurrence graph is created in the following way: we start with an empty graph and for every news article we create a subgraph with all the companies as nodes, and with every two nodes connected by an edge (a complete graph). For instance, for the article A_4: $\{C_1, C_2, C_3, C_4\}$ we create the following list of edges: $\{\{C_1, C_2\}, \{C_1, C_3\}, \{C_1, C_4\}, \{C_2, C_3\}, \{C_2, C_4\}, \{C_3, C_4\}\}$, with the edge weight of 1. Every subsequent article adds a new subgraph to the whole graph. Note that the multiple mentions of a single company in the same article text are discarded and does not influence the resulting network.

In order to include the temporal aspect of the co-occurrence data, each edge is assigned a date that equals the publication date of article A. This allows us to obtain a number of co-occurrence between company pairs in the desired timeframe by calculating the weight using temporal function t. In order to calculate the weight $w(e)$ of edge e between two vertices C_A and C_B we sum the number of edges $\{C_A, C_B\}$, where the value of temporal function $t(e)$ is between date d_1 and d_2.

Figure 2 presents an overview of the co-occurrence graph for the period of Q1 2013. The graph has been drawn using Gephi[2] software and Fruchterman Rheingold graph layout [13]. Darker colors signify higher degree (nodes) or higher weight (edges). The node and edge labels have been removed for brevity.

[2] Gephi: The Graph Visualisation Platform https://gephi.org/

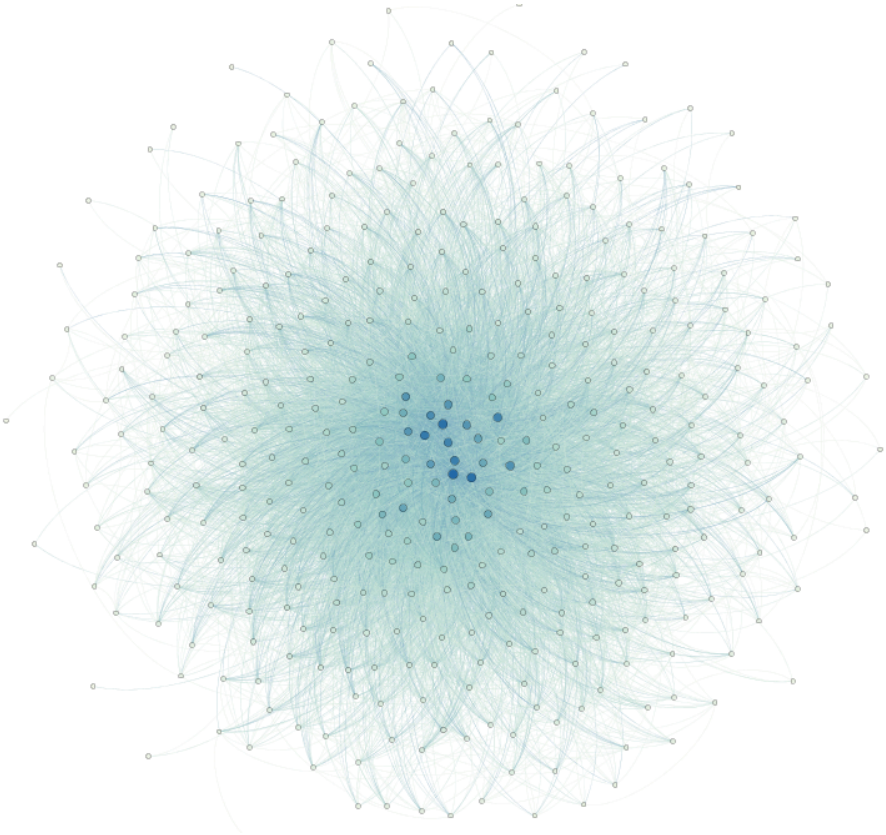


Figure 2: The graph of co-occurrence for Q1 2013

The most important aspect to observe is that the graph density is very high, accounting for high number of relation between companies. Also there is huge concentration of co-occurrence between few companies in the middle of the graph (dark blue dots), which account for stocks, which are most popular in the news, such as: MSFT, AAPL, GOOGL, AMZN.

4 Prediction Model

4.1 Company Relationship Coefficient

Based on the company relation network graph described in Section 3 we develop a Company Relationship Coefficient CRC. The CRC value describes the amount of co-occurrence of company pair $\{C_A, C_B\}$ as an edge weight $w(e)$ in the graph between node C_A and C_B and $t(e)$ is in the time period (d_1, d_2). The CRC value is normalized using the sum of all edge weights in the graph in a given timeframe in order to facilitate comparing CRC values across various time periods. The CRC value between

companies is higher when both companies appear more often in news articles or zero when there is no co-occurrence.

The Relationship coefficient is further used as a predictor variable in the process of forecasting using the linear model.

4.2 Time Series correlation

There are numerous methods for estimating financial correlation between two stocks. Our work is based on commonly used daily returns measure. We calculate it as a 1-day percentage change within the stock price time series (taking the daily closing stock prices):

$$R(t) = \frac{S(t_2) - S(t_1)}{S(t_1)}$$

As a result we obtain the vector of daily returns for the chosen stock. We compute the Stock Correlation Coefficient (SCC) for the company pair as a correlation coefficient of two vectors of daily returns using the Pearson's formula. The SCC value is higher when both stocks move up or down often at the same time. High SCC value also means that both stocks are generating either profit or loss at the same time, resulting in unbalanced portfolio and higher investment risk.

The stock prices time series have been obtained through Yahoo Finance API by using the YQL interface[3].

4.3 Problem Statement

In this paper we analyze the hypothesis *H* that the co-occurrence of the companies in news has a positive effect on the correlation of daily returns between company pairs. To verify it, we use linear regression with least squares estimation to observe if there is significant relation between CRC and SCC variables.

In this case the predictor variable *x* is the amount of co-occurrences between two companies (CRC value) in a given timeframe and the observed result *y* is the correlation coefficient of daily returns for the same company pair (SCC value) and the same timeframe. The null hypothesis H_0 is that there is no effect of variable *x* on variable *y*. In order to analyze if the model is suitable for predictions, we use the P-value measure with the threshold of 0.05. We assume that for P-values < 0.05 the null hypothesis can be rejected and that there is statistical evidence that our predictor variable has an effect on the observed daily returns correlations.

To give better over we additionally condition the model according to the industry sectors. Therefore all the companies were grouped according to the GICS taxonomy[4] into the following sectors: "Telecommunication Services", "Utilities", "Health Care",

[3] http://developer.yahoo.com/yql

[4] The Global Industry Classification Standard (GICS®) http://www.msci.com/products/indexes/sector/gics/

"Industrials", "Information Technology", "Materials", "Consumer Discretionary", "Consumer Staples", "Energy", "Financials". When two companies belong to different sectors, it is assigned the group label "Mixed".

4.4 Evaluation Results

The evaluation has been carried in two steps: (i) first a regression analysis has been performed between CRC and SCC variables separately for each industry sector and for quarterly time periods (ii) then we calculated the P-values for each regression to see how probable is the hypothesis *H*. The linear regression has been calculated by taking all company pairs from the co-occurrence graph with weight > 1 and calculating both CRC (normalized weight) and SCC (based on daily returns correlation). Each CRC (x variable) and SCC (y variable) pair is one observation. Figure 3 shows the regression plots for Q1 2013, with observation points in blue and fitted regression line in red. The x-axis is log-transformed.

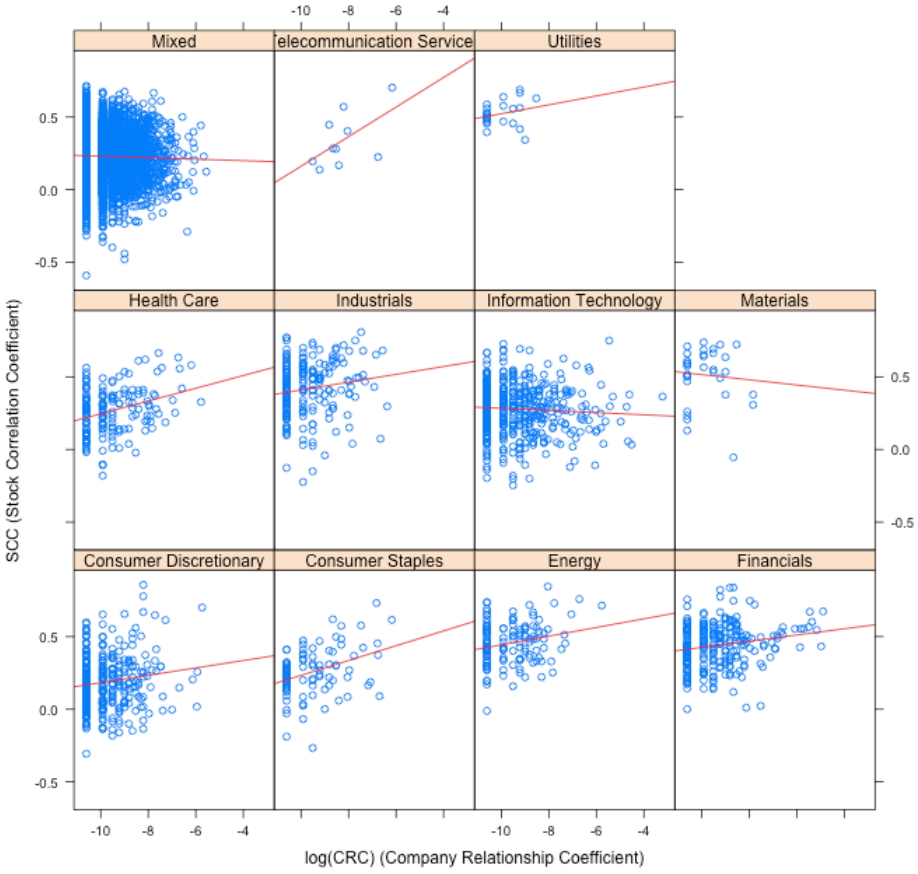


Figure 3: Regression chart for Q1 2013

Observing Q1 2013 data we can already notice a clear trend for the following sectors: “Telecommunication Service”, “Health Care”, “Industrials”, “Consumer Staples”, “Consumer Discretionary”, “Energy” and “Financials”. For those sectors, the CRC variable can serve as a predictor for SCC. On the other hand, the “Information Technology” or “Mixed” sectors doesn’t provide any significant insight on CRC/SCC relations due to a lot of noise.

To better estimate the statistical significance, we performed calculation of P-values separately for each quarter to see if they hold over time. The result is presented in Table 1. Cells highlighted in green mark the quarters where the P-value is below the threshold of 0.05.

	P-value				
Sector	**Q1**	**Q2**	**Q3**	**Q4**	**P-value < 0.05**
Health Care	0,0004	0,0072	0,0014	0,0116	100%
Consumer Staples	0,0081	0,2114	0,0024	0,0077	75%
Energy	0,0009	0,0266	0,0678	0,1424	50%
Financials	0,0075	0,1467	0,0002	0,0988	50%
Utilities	0,6483	0,2641	0,0330	0,0349	50%

Table 1: Evaluating the regression model

The most significant results were obtained for the “Health Care” and “Consumer Staples” industrial sectors, where the CRC/SCC relation is strong and holds for most of the time. For “Energy”, “Financials” and “Utilities” the strong trend appears only in two out of four quarters. The remaining sectors have been filtered out, as they did not produce significant results for more than one quarter.

5 Conclusions and Future Work

This article presented an approach for predicting stock return correlations based on the unstructured data in the form of financial news. The experiment was conducted for historical data from year 2013 and demonstrated that for certain industrial sectors we were able to prove significant relation between co-occurrence of companies in the news articles and correlation of daily returns of their stocks.

The results of this experiment shows that analyzing very noisy data, such as automatically extracted information based on the news articles, can still lead to insightful discoveries. In our case, observing very simple company relationships can predict in certain cases the stock correlation. This can result in better portfolio optimization and new ways of mitigating investment risk.

As a future work, we are working on generating a new dataset of financial news and improving the company detection through a more accurate NER (Named Entity Recognition) process and knowledge-based information extraction. We are also working on analyzing why some industrial sectors work better than the others and what are the sources of noise in the data in order to understand how to seek better accuracy in the unstructured data sources.

8 Acknowledgments

This work was supported by the Spanish Ministry of Science and Innovation under the project FLORA (TIN2011-27405).

References

1. F. Gens, “IDC Predictions 2012: Competing for 2020”, http://cdn.idc.com/research/Predictions12/Main/downloads/IDCTOP10Predictions2012.pdf (accessed March 2, 2014).
2. ScienceDaily, "Big Data, for better or worse: 90% of world's data generated over last two years." http://www.sciencedaily.com/releases/2013/05/130522085217.htm (accessed March 2, 2014).
3. J. Smailović, M. Žnidaršič, and M. Grčar, “Web-based experimental platform for sentiment analysis.”
4. J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič, “Predictive Sentiment Analysis of Tweets: A Stock Market Application,” in *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, Springer Berlin Heidelberg, 2013, pp. 77–88.
5. A. Klein, O. Altuntas, T. Hausser, and W. Kessler, “Extracting Investor Sentiment from Weblog Texts: A Knowledge-based Approach,” in *Commerce and Enterprise Computing (CEC), 2011 IEEE 13th Conference on*, 2011, pp. 1–9.
6. M. Saveski and M. Grčar, “Web Services for Stream Mining: A Stream-Based Active Learning Use Case,” *eCML PKDD 2011*, p. 36, 2011.
7. G. Rizzo, “Nerd: evaluating named entity recognition tools in the web of data,” 2011.
8. B. Sluban and M. Grčar, “URL Tree: Efficient Unsupervised Content Extraction from Streams of Web Documents,” in *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, 2013, pp. 2267–2272.
9. T. Von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. W. Fellner, “Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges,” in *Computer graphics forum*, 2011, vol. 30, no. 6, pp. 1719–1749.
10. Z. Ma, O. R. L. Sheng, and G. Pant, “Discovering company revenue relations from news: A network approach,” *Decis. Support Syst.*, vol. 47, no. 4, pp. 408–414, Nov. 2009.
11. B. Sluban, M. Grcar: Efficient Unsupervised Content Extraction from Streams of Web Documents. In: Proceedings of the 22nd International Conference on Information and Knowledge Management (CIKM 2013). Burlingame, California, USA, 2013.
12. M. Grčar, P. Kralj, V. Dinev, A. Klein: "D3.2: Ontology reuse and evolution" FIRST: Large scale information extraction and integration infrastructure for supporting financial decision making. (September 2012)
13. Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph Drawing by Force-Directed Placement. Software: Practice and Experience, 21(11)

Appendices

1. **WaSABi 2014: Keynote Slides: Crossing the Chasm with Semantic Technologies Marin Dimitrov**

2. **WaSABi 2014: Slides: Breakout Brainstorming Session Group 1 Breakout Group 1**

3. **WaSABi 2014: Slides: Breakout Brainstorming Session Group 2 Breakout Group 2**

Crossing the Chasm with Semantic Technology

Marin Dimitrov (CTO)

WaSABi'2014

Contents

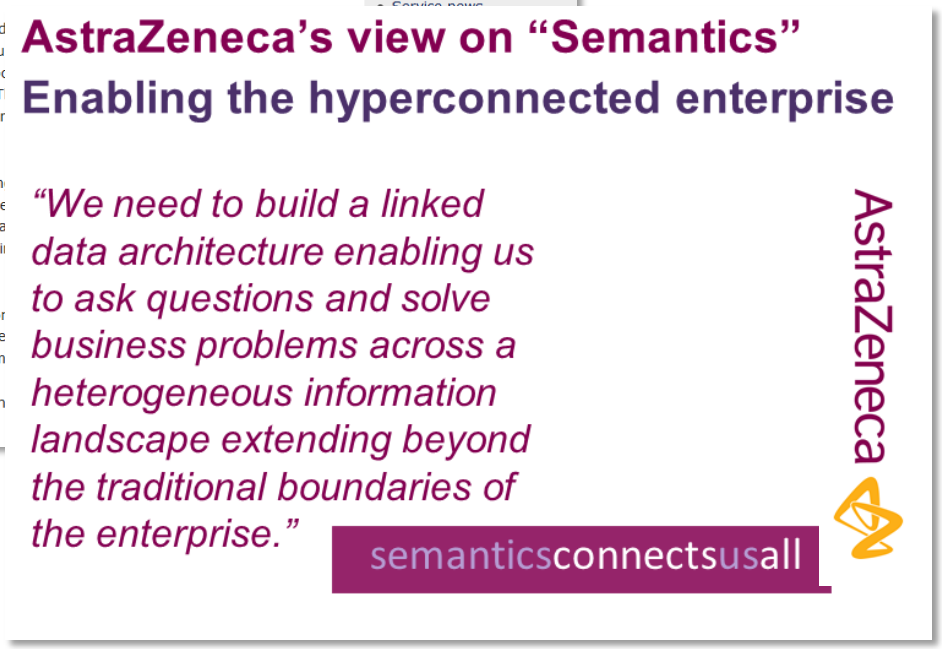
- Semantic Technology – Fad or Future?
- Innovation, Hype & Chasm
- Lessons Learned

About Ontotext

- Provides products & solutions for content enrichment and metadata management
- Major clients and industries
  - Media & Publishing
  - Health Care & Life Sciences
  - Cultural Heritage & Digital Libraries
  - Government
  - Recruitment

SEMANTIC TECHNOLOGY – FAD OR FUTURE?

Life Sciences



Cultural Heritage

LINKED DATA

The Importance of the Semantic Web To Our Cultural Heritage

By Jennifer Zaino on May 20, 2014 9:00 AM

Earlier this year The Semantic Web Blog reported that the Getty Research Institute has released the Art & Architecture Thesaurus (AAT) as Linked Open Data. One of the external advisors to its work was Vladimir Alexiev, who leads the Data and Ontology Management group at Ontotext and works on many projects related to cultural heritage.

Ontotext's OWLIM family of semantic repositories supports large-scale knowledge bases of rich semantic information, and powerful reasoning. The company, for example, did the first working implementation of CIDOC CRM search; CIDOC CRM is one of these rich ontologies for cultural heritage.

We caught up with Alexiev recently to gain some insight into semantic technology's role in representing the cultural heritage sphere. Here are some of his thoughts about why it's important

Publishing

BBC | Sign in | News | Sport | Weather | Shop | Capital

INTERNET BLOG

Previous | Home | Next

Linked Data: Connecting together the BBC's Online Content

Tuesday 19 February 2013, 09:31

Oliver Bartlett
Product Manager

COMMENTS (11)

Tagged with: Linked Data, BBC News

Hi I'm Oli Bartlett, product manager for the BBC's Linked Data Platform.

The Linked Data Platform is one of the legacies of the BBC Sport **2012 Olympi**
have read my **blog post** on the work we did for the **Olympic Data Service**.

One aspect of the service delivered the semantic framework for the 10,000 athle
event, discipline, country and venue.

This framework provides the semantic graph of data (the **linked data** containing
venues and their associations with each other) and the **API**s on this data.

BBC | Sign in | News | Sport | Weather | Shop | Capital

INTERNET BLOG

Previous | Home | Next

Linked Data: new ontologies website

Wednesday 30 April 2014, 10:05

Sofia Angeletou
Data Architect

COMMENTS (6)

Share

Tagged with: BBC Online, Linked Data

Hello, I'm Sofia Angeletou and I'm the Data Architect for the Linked Data Platform (LDP), which builds the BBC's services for creating and publishing linked data.

I'm going to talk to you about our new **/ontologies** site which we released last week and where you can find the **ontologies** that BBC uses to support **BBC Sport, Education**, news prototypes and soon BBC **Music** and Radio programmes.

What is it and why are we doing it?

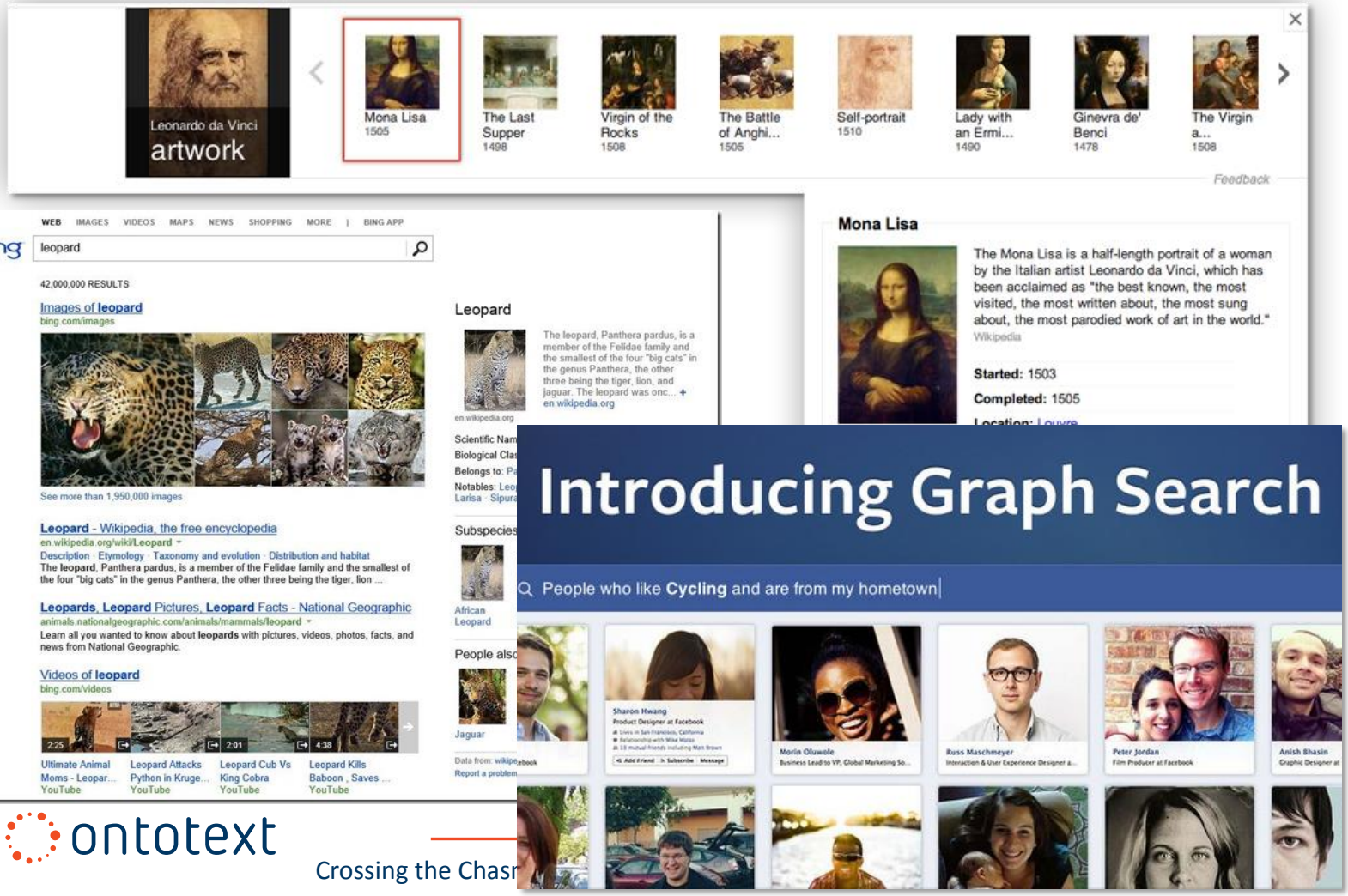
Oli Bartlett, the owner of the Linked Data Platform, has **explained** how we have expanded the reach of linked data within the BBC to more audience facing products and presented our ambitions to using linked data as glue for the plethora of content the BBC produces. As a direct result of this, more models are being built to support additional functionality and cover new and diverse domains of interest.

bbc.co.uk/ontologies is a human friendly view of the data models in the Linked Data Platform and is meant to give a comprehensive understanding of which ontologies the BBC uses, why and how. This is provided for members of the public and anyone who wants to get a better understanding of the BBC's Linked Data.

Showcases & Talks at SemTechBiz Since 2010

- Accenture, AFP, Alcatel-Lucent, Autodesk, BestBuy, Boeing, CapGemini, Cisco, Daimler, Disney Media, DoD, eBay, EC, Elsevier, EMC, Fujitsu, Gartner, Getty, Google, IBM, Library of Congress, Lockheed Martin, Mayo Clinic, Merck, Microsoft, NASA, Novartis, Oracle, Press Association, Renault, Salesforce, SAP, Siemens, Statoil, Teradata, TIBCO, Walmart, Wells Fargo, Yahoo, Yandex

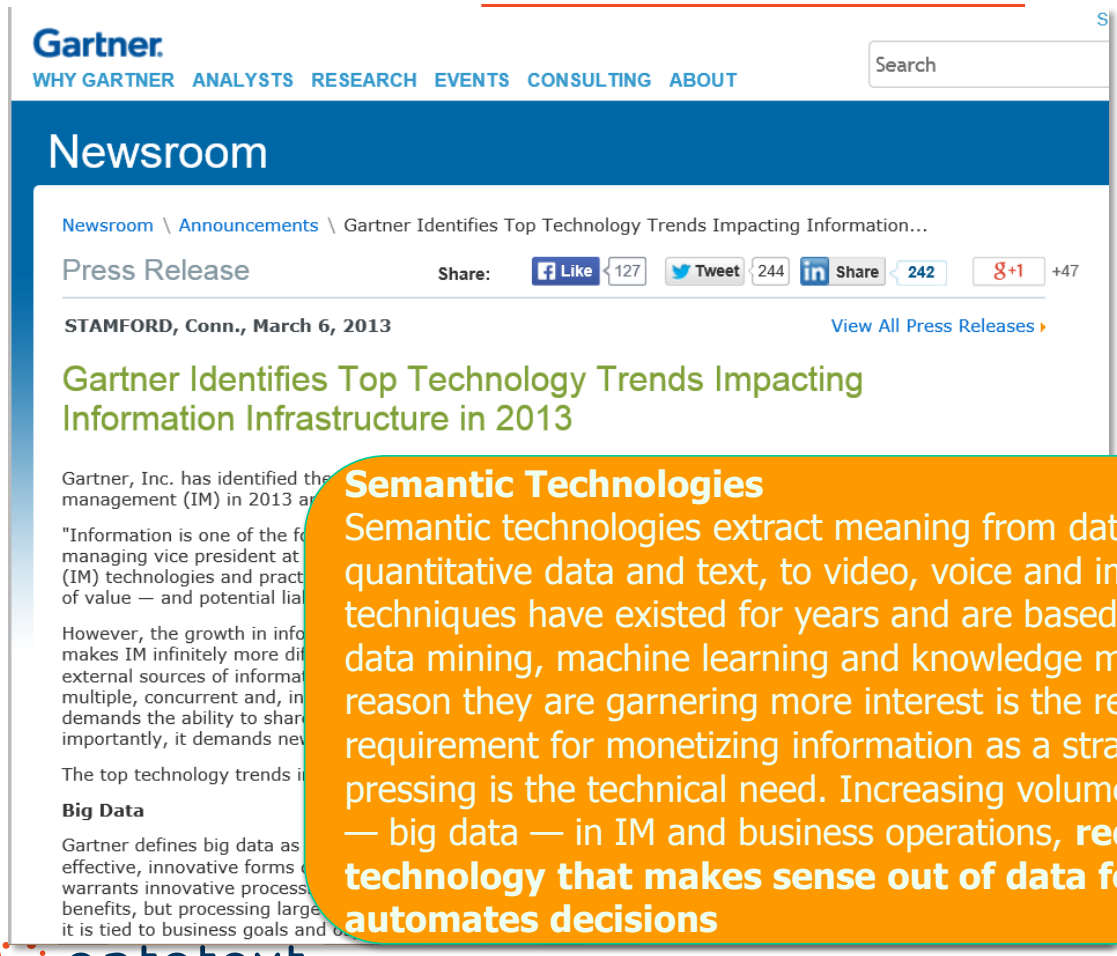
Knowledge Graphs



ontotext

Crossing the Chasm

Top Information Management Trends 2013 (Gartner)



Semantic Technologies

Semantic technologies extract meaning from data, ranging from quantitative data and text, to video, voice and images. Many of these techniques have existed for years and are based on advanced statistics, data mining, machine learning and knowledge management. One reason they are garnering more interest is the renewed business requirement for monetizing information as a strategic asset. Even more pressing is the technical need. Increasing volumes, variety and velocity — big data — in IM and business operations, **requires semantic technology that makes sense out of data for humans, or automates decisions**

A Different Point of View

InformationWeek CONNECTING THE BUSINESS TECHNOLOGY COMMUNITY

Sea

Home News & Commentary Authors Slideshows Video Reports White Papers Events Interop

STRATEGIC CIO | SOFTWARE | SECURITY | CLOUD | MOBILE | BIG DATA | INFRASTRUCTURE | GOVERNMENT

SOFTWARE // INFORMATION MANAGEMENT

COMMENTARY
1/7/2014
09:06 AM

Seth Grimes
Commentary

Connect Directly

17 COMMENTS
COMMENT NOW

Login

Semantic Web Business: Going Nowhere Slowly

The semantic web vision persists, but the tools and processes don't stand up to today's data chaos.

I've been a semantic web skeptic for years. SemWeb is a narrowly purposed replica of a subset of the World Wide Web. It's useful for information enrichment in certain domains, via a circumscribed set of tools. However, the SemWeb offers a vanishingly small benefit to the vast majority of businesses. The vision persists but is unachievable; the business reality of SemWeb is going pretty much nowhere.

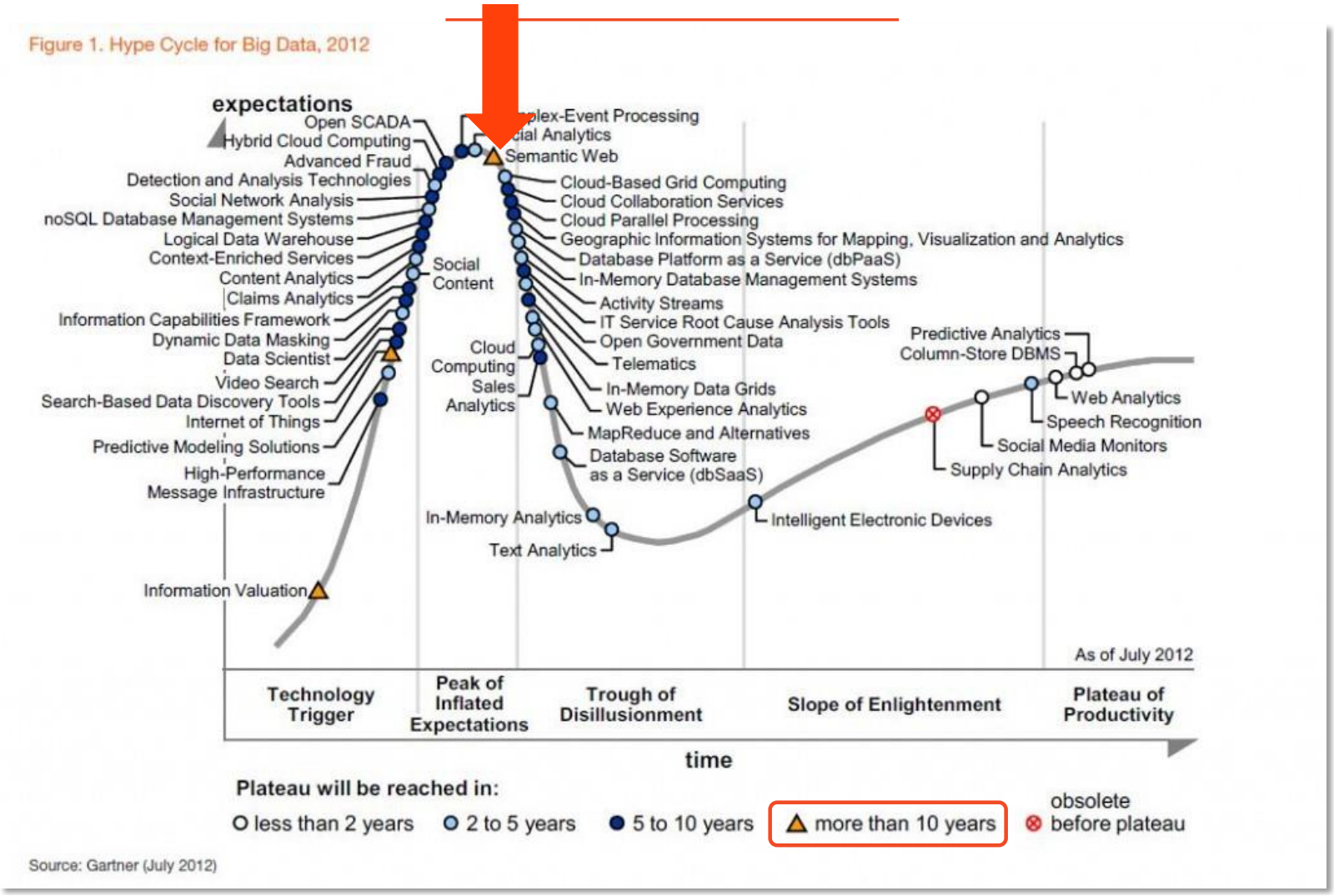
The SemWeb dream centers on sharing linked data via the W3C's Resource Description Framework protocol. There is no question that SemWeb aspires to a worthy goal, but its tools and processes are no match for the reality of never-diminishing online, social, and enterprise data chaos. SemWeb can't keep up with the flow, even on the limited portion of the data universe that is published on the World Wide Web. We will never achieve its ideal universe of

REPORTS

The Agile Ar

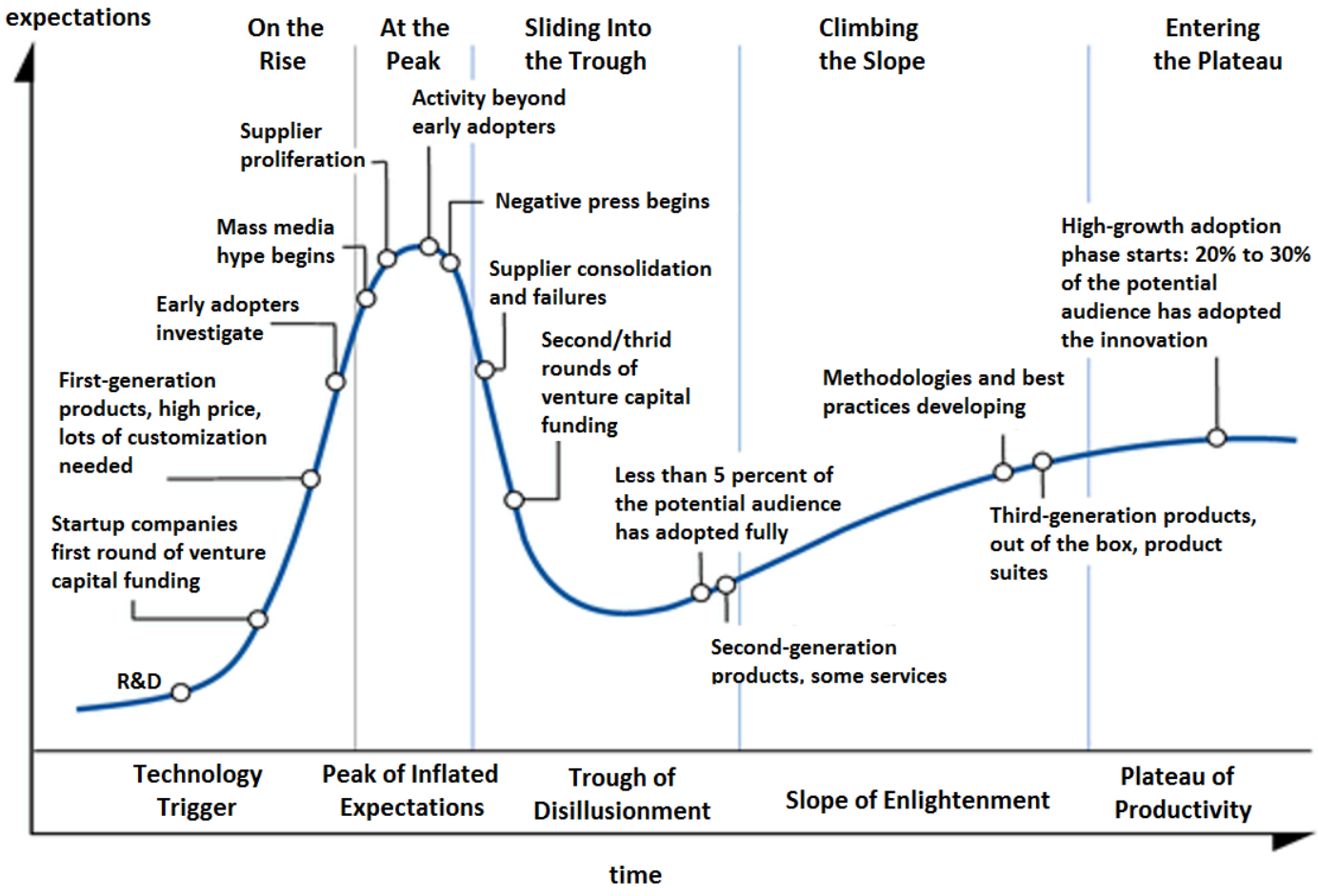
When it comes to ma
archiving systems as
well-designed archiv
restores, ease searc

Big Data Hype Cycle 2012 (Gartner)

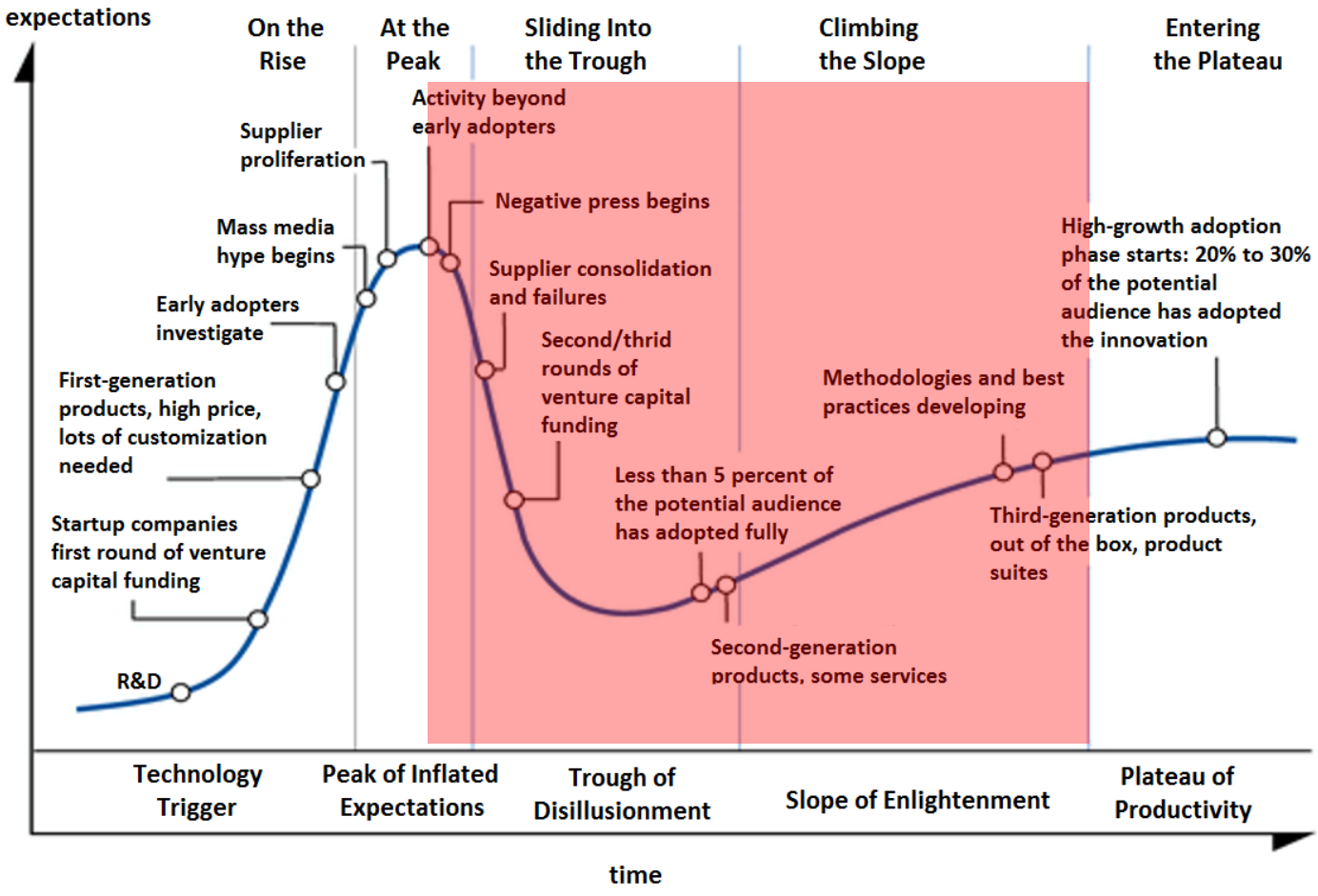


INNOVATION, HYPE & CHASM

Technology Hype Cycle (Gartner)



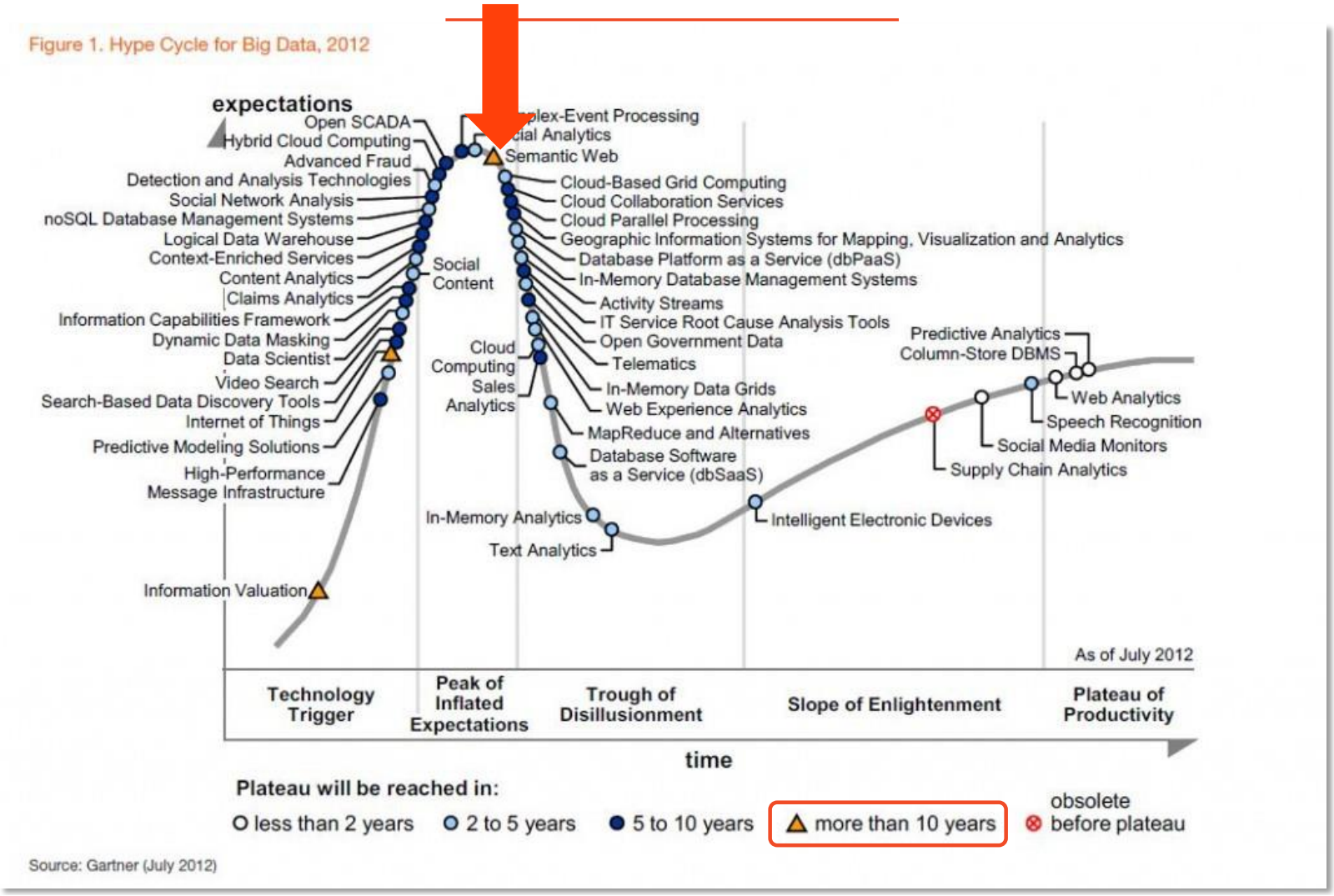
Time-to-value Gap (Gartner)



Time-to-value Gap (Gartner)

- **Performance**
  - Consistent reliability, availability, quality
- **Integration**
  - Innovation must fit into existing environments & constraints
- **Penetration**
  - Critical mass of adopters required
- **Payback**
  - Deriving business values, cost savings, ROI
  - *Amounts* and/or *timing* usually difficult to estimate

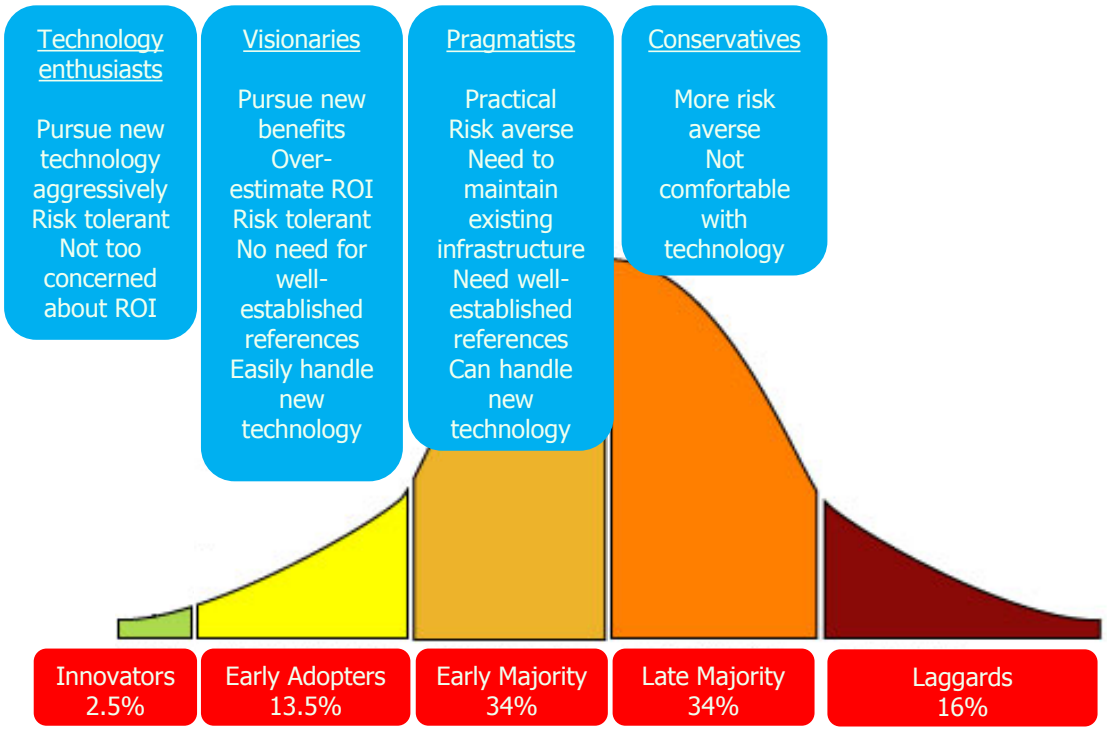
Big Data Hype Cycle 2012 (Gartner)



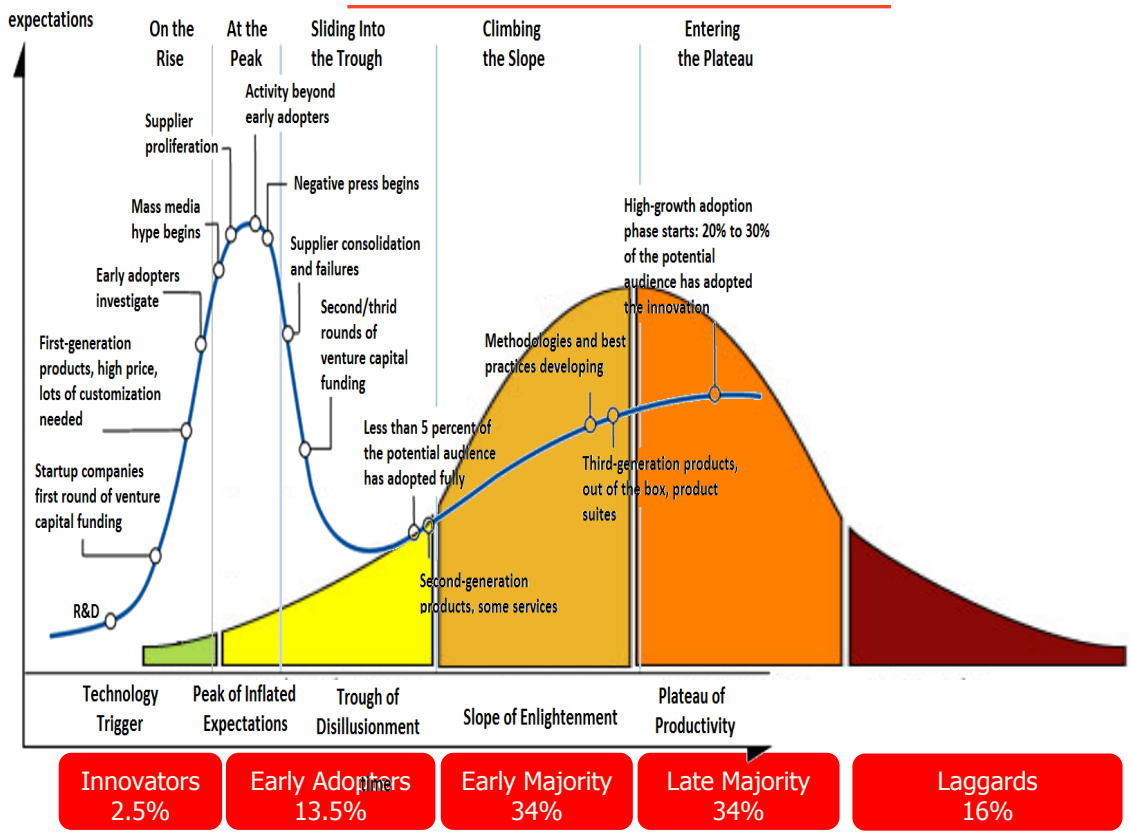
Semantic Technologies on the Hype Cycles 2013 (Gartner)

- Hype cycles for Web Computing, Information Infrastructure, Enterprise Information management, etc.
- Related technologies
  - Graph databases, Semantic Web, metadata management, content/text analytics, taxonomy & ontology management, entity resolution & analysis
  - Positioned in the early phases: on the rise / at the peak / sliding into the trough

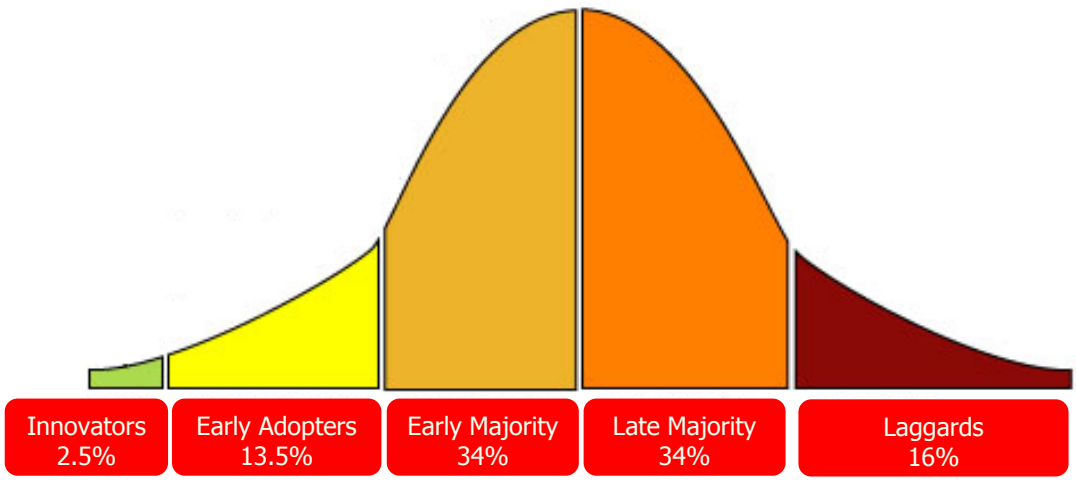
Technology Adoption Lifecycle



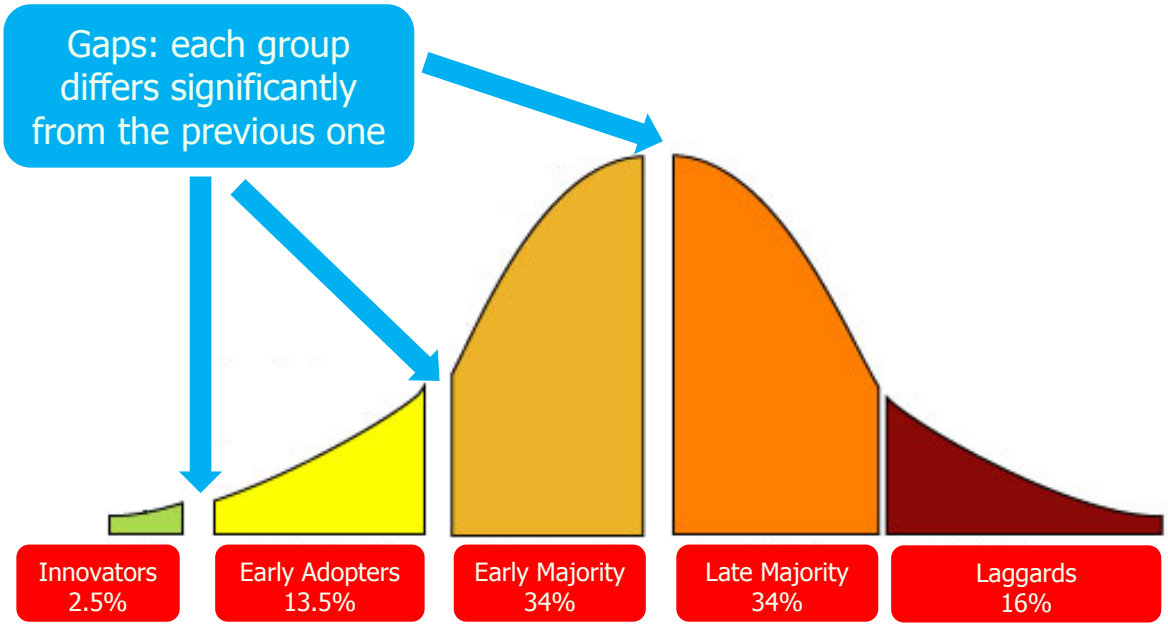
Technology Adoption Lifecycle & Gartner Hype Cycle



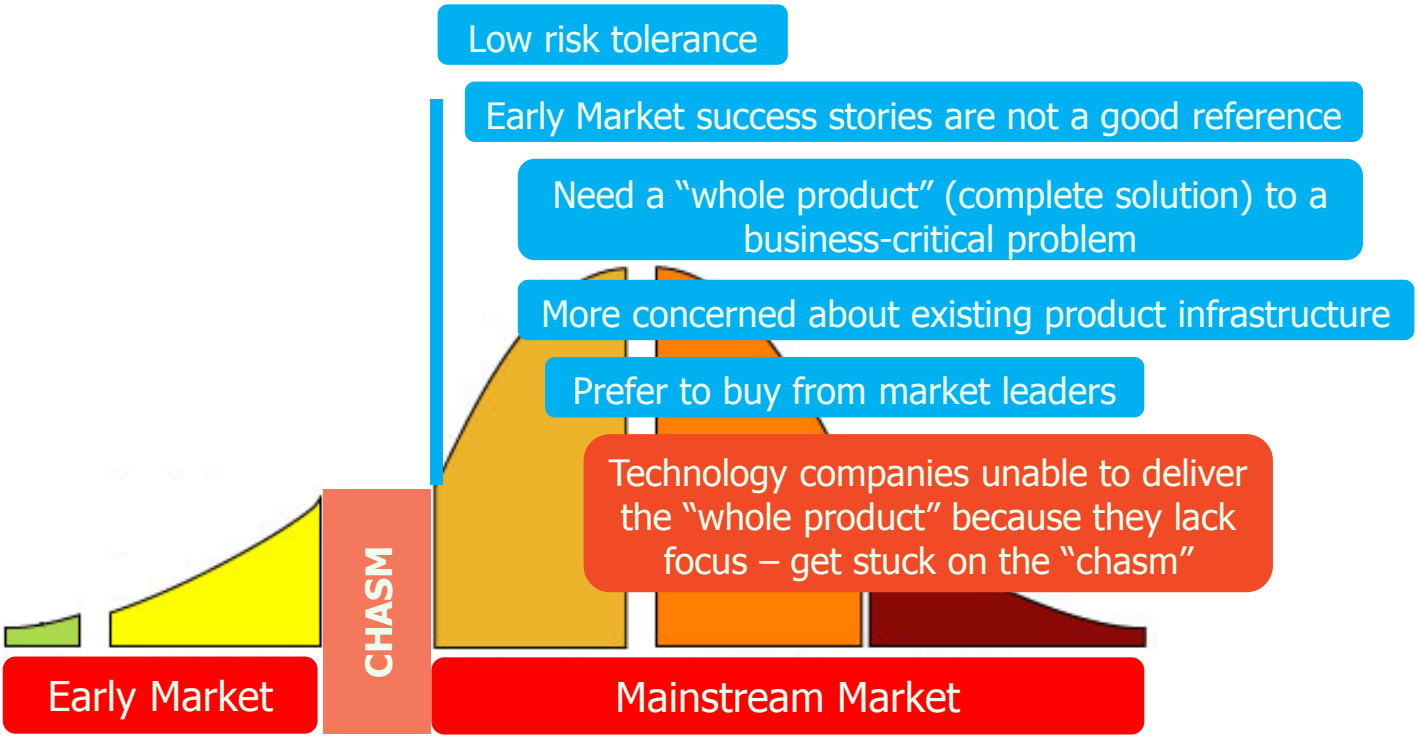
Gaps in the Technology Adoption Lifecycle



The Chasm (Geoffrey Moore)

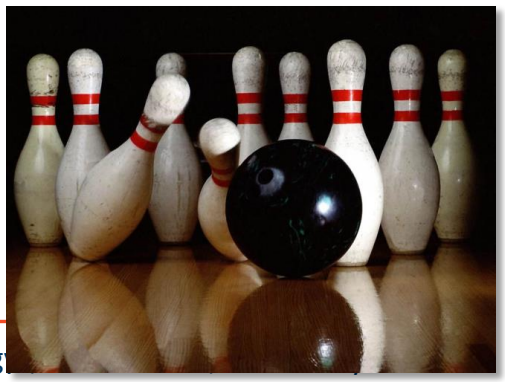


The Chasm (Geoffrey Moore)

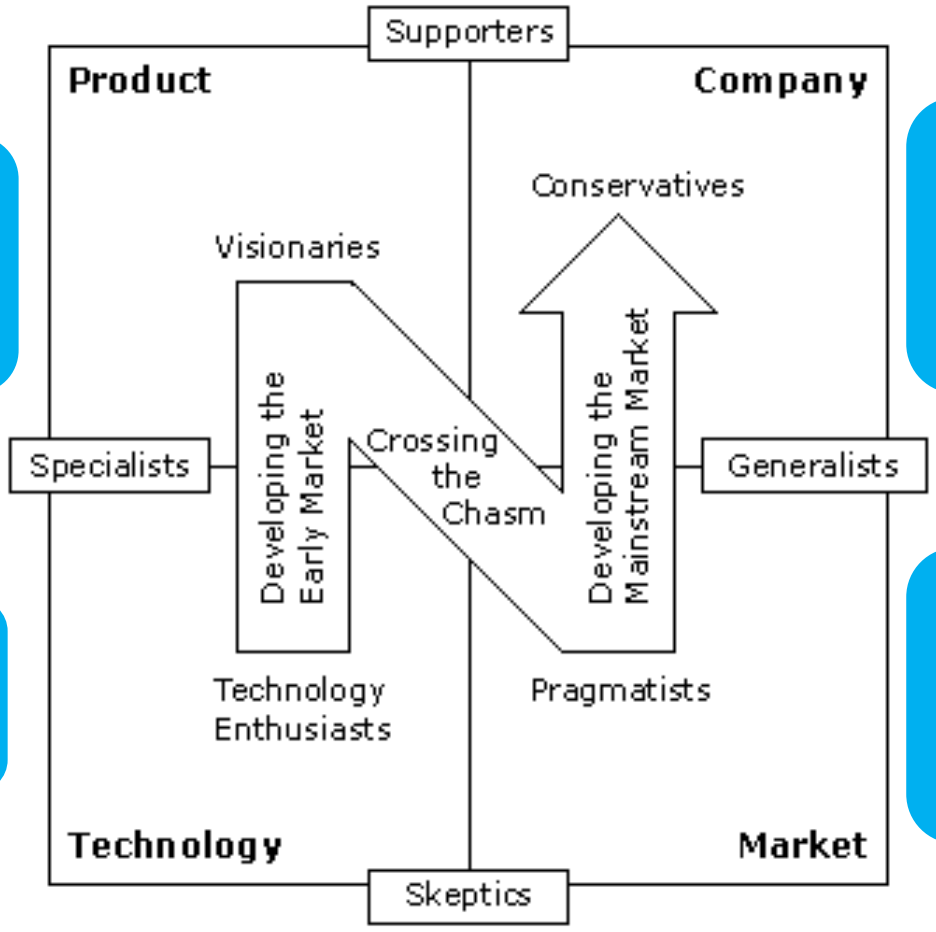


Crossing The Chasm (Geoffrey Moore)

- Identify one attractive mainstream market customer (niche)
- Focus on providing the "whole product" for their problem
  - Partnerships with other providers may be required
  - A reference "success story" for other mainstream buyers
- Become the market leader in the niche & move into adjacent niches
  - Bowling alley effect



The Competitive-Positioning Compass (Geoffrey Moore)



Benchmarks
Product reviews
Sales volume
Trade press
Visionary endorsement

Revenues
Partners
Top customers
Full product line
Business press
Financial analyst endorsements

Architecture
Trials & demos
Technology press
Guru endorsement

Market share
Standards & certification
3rd party support
Industry press
Industry analyst endorsements

ontotext

LESSONS LEARNED

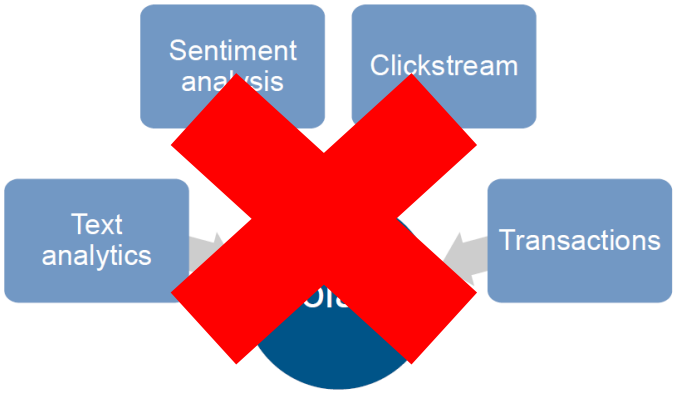
Lessons Learned

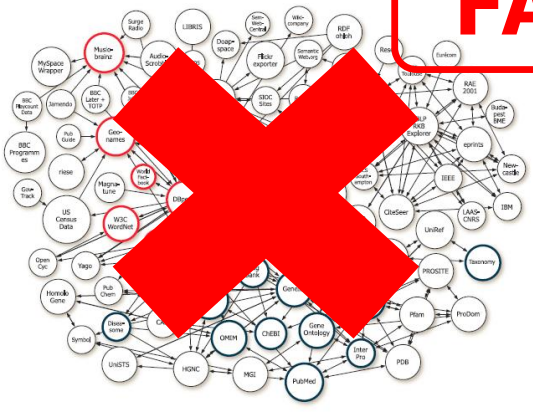
- Innovations go through ups and downs before reaching the productivity phase
  - Customer: experimentation and patience often required before value is delivered
  - Customer: TCO often higher than expected
  - Provider: target the value gaps early: *Performance, Integration, Penetration, Payback*
- Understand the technology adoption challenges
  - Early market success *does not* translate to mainstream market success
  - Different strategies for delivering value to Enthusiasts, Visionaries, Pragmatists & Conservatives

Lessons Learned

- Follow the "chasm crossing" principles
  - Focus on an attractive mainstream customer / niche
  - Find partners & deliver the "whole product" (complete solution) that solves a business-critical problem
  - Use the success story as a 1st reference point
  - Move into adjacent niches (bowling alley effect)
- Clearly convey the benefits of your solution
  - *Not* via a product feature list or benchmarks
  - Speak the language of the customer
  - How is your solution better than the current one?
  - Measurable returns and timeframe for achieving them

Clearly Convey the Benefits of Your Solution

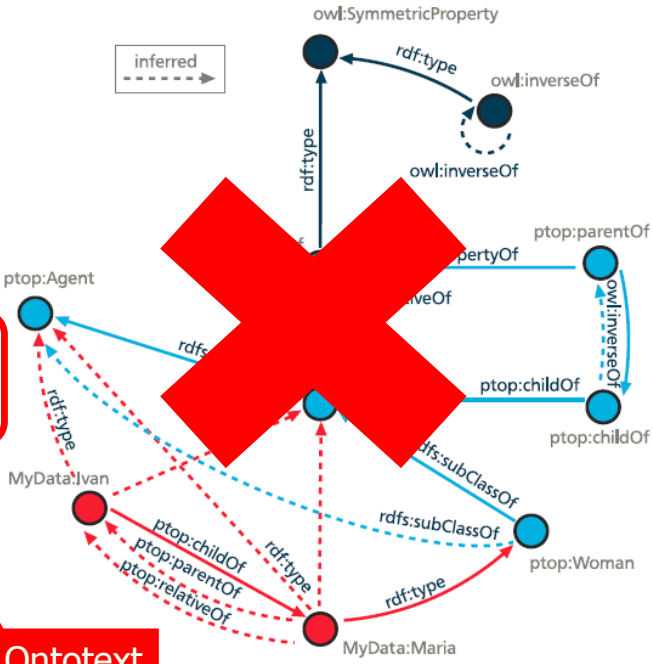




Gartner

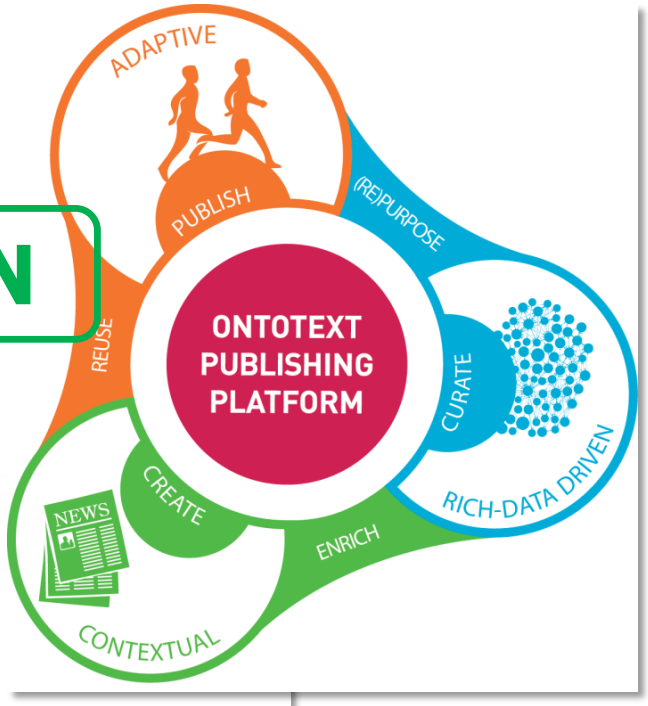
FAIL

Ontotext brochure 2012



Clearly Convey the Benefits of Your Solution

WIN

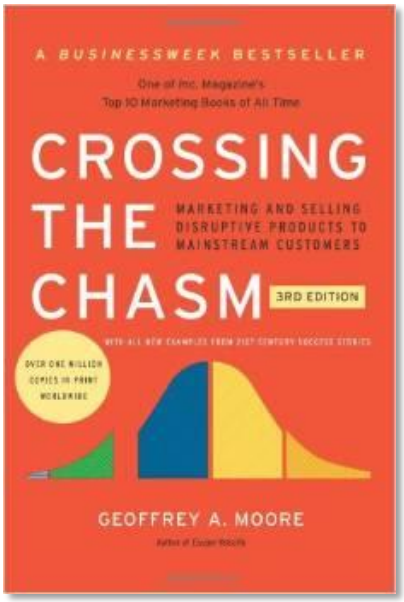


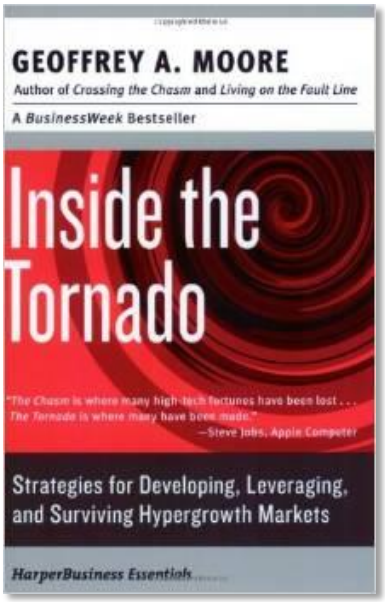
How can you use our database to gain competitive advantage?

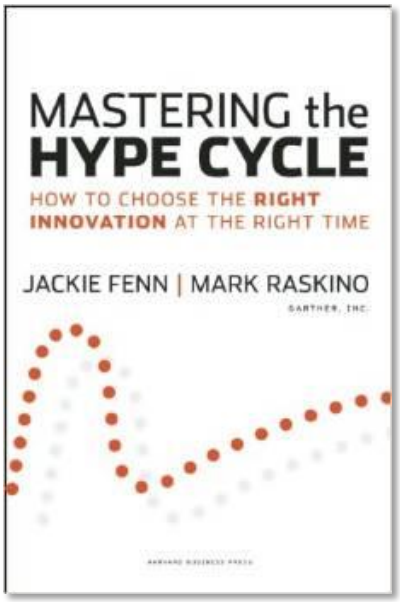
1. Which are the most active Job advertisers?
2. Which are the agencies and employers that do not advertise on your Job board?
3. Which is the right Job board for your industry sector and a region?

Contact us to find out more http://www.ontotext.com/ukjobsdata

Recommended Reading







Q & A

Thank you!

Breakout Brainstorming Session

Group 1

WaSABi 2014: Workshop on Semantic Web Enterprise Adoption and Best Practice

2014-05-26

Adoption Challenges

- Longevity of URIs in a changing environment.
- Changes of vocabularies/datasets while keeping the same namespace can cause applications to fail.
- How do we (and particularly industry!) know which vocabularies and technologies that have strong community backing?

Longevity of URIs in a Changing Environment

- Newly developed vocabularies believed to reach external uptake should be hosted by W3C, or possibly PURL.org.
- For existing vocabularies; log and publish statistics on read access so that it can be determined which existing vocabularies are heavily used, in order to learn how significant this potential problem is.
- Possibly consider how BitTorrent anchor links work; could their approach to DNS-less resource resolution work also for Semantic Web data?

Changes to Vocabularies

- **Never** change a published (i.e., available via a publicly resolvable URI) vocabulary or dataset - only add new URI:s with new versioning info.
- Further work needed on Ontology Evolution and logging, particularly regarding presenting change deltas in an accessible user-friendly or at least developer-friendly way.
- Related to the above: the use of the deprecation warning and versioning annotations in OWL needs to be increased.
- Could a system be developed to generate reports from a codebase, detailing which datasets/vocabularies/libraries that the code depends upon?

Communicating Future-Proof Artefacts

- Emphasize the use of and development of indices like LOV and LodStats, possibly also for other types of artefacts than just vocabularies.
- Could a social networking-based community portal be of use? Currently the semantic web research community exists across a great deal of sites, keeping it all in one place and combining this with a user-maintained directory of quality components and technologies might be beneficial.

Breakout Brainstorming Session

Group 2

WaSABi 2014: Workshop on Semantic Web Enterprise Adoption and Best Practice

2014-05-26

Adoption Problems

- Difficult to train developers, steep learning curve.
- Basic developer problems:
  - ORM/DAO transformation
  - ETL
  - LDP/Middleware
- Where to make a cut skipping SW internals?
- Scalability: Volume, Velocity, Variety (Veracity, Variability, Visualization, Value)

Adoption Solutions

- New abstraction layers (less functionalities).
  - REST-ful APIs.
  - Black boxing the technology not always possible.
- Research directions are changing towards application.
  - Companies in charge (?)