Semantic Annotation and Linking of Competitive Intelligence Reports for Business Clusters

[Position Paper]

Jan Nemrava
Dept. Information and Knowledge Engineering
University of Economics, Prague
Winston Churchill Sq. 4
Prague, Czech Republic
nemrava@vse.cz

Tomáš Kliegr
Dept. Information and Knowledge Engineering
University of Economics, Prague
Winston Churchill Sq. 4
Prague, Czech Republic
tomas.kliegr@vse.cz

Vojtěch Svátek
Dept. Information and Knowledge Engineering
University of Economics, Prague
Winston Churchill Sq. 4
Prague, Czech Republic
svatek@vse.cz

Martin Ralbovský
Dept. Information and Knowledge Engineering
University of Economics, Prague
Winston Churchill Sq. 4
Prague, Czech Republic
ralbovsm@vse.cz

Jiří Šplíchal
Tovek spol. s r.o.
Chrudimská 1418/2
Prague, Czech Republic
splichal@tovek.cz

Tomáš Vejlupek
Tovek spol. s r.o.
Chrudimská 1418/2
Prague, Czech Republic
vejlupek@tovek.cz

ABSTRACT

Competitive intelligence (CI) is a sub-discipline of business intelligence that supports the decision makers in understanding the competitive environment by means of textual reports prepared based on public resources. CI is particularly demanding in the context of larger business clusters. We report on a long-term project featuring large-scale manual semantic annotation of CI reports wrt. business clusters in several industries. The underlying ontologies are the result of collaborative editing by multiple student teams. The results of annotation are finally merged into CI maps that allow easy access to both the original documents and the knowledge structures.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; I.7 [**DOCUMENT AND TEXT PROCESSING**]: Document Management

General Terms

Collaborative annotation, Ontology Engineering

OBI'08, October , 2008, Karlsruhe, Germany.

Keywords

Competitive Intelligence reports, document annotation, shared ontology population

1. INTRODUCTION

Business Intelligence (BI) is still, by many, primarily perceived as the collection of activities related to analysing data from various systems internal to an organisation. Such data reflect the processes taking place within the organisation; as the main goal of BI is considered the optimalisation of such processes, i.e. 'how to do things properly'. However, in order to know 'the proper things to do' in a competitive environment, there is a strong need for information from the organisation's surroundings, which can be obtained either from external resources or from employees who are familiar with these surroundings. The process of collecting, analysing and presenting (to the management) such information has recently been labeled as *competitive intelligence* (CI). In general, CI is an ethical business discipline that supports decision makers in understanding the competitive environment. Its main vehicle are *CI reports*, which are prepared on the basis of open sources such as web pages, articles or business registries.

A *business cluster* is a geographic concentration of interconnected businesses, suppliers, and associated institutions in a particular field.[1] CI efforts within clusters are more complicated than those within individual companies, as different cluster members may perceive the market situation differently and also establish liaisons with other industries in different ways. On the other hand, the cost of CI can be shared across the cluster, assuming the benefits of exploit-

[1] `http://en.wikipedia.org/wiki/Business_cluster`

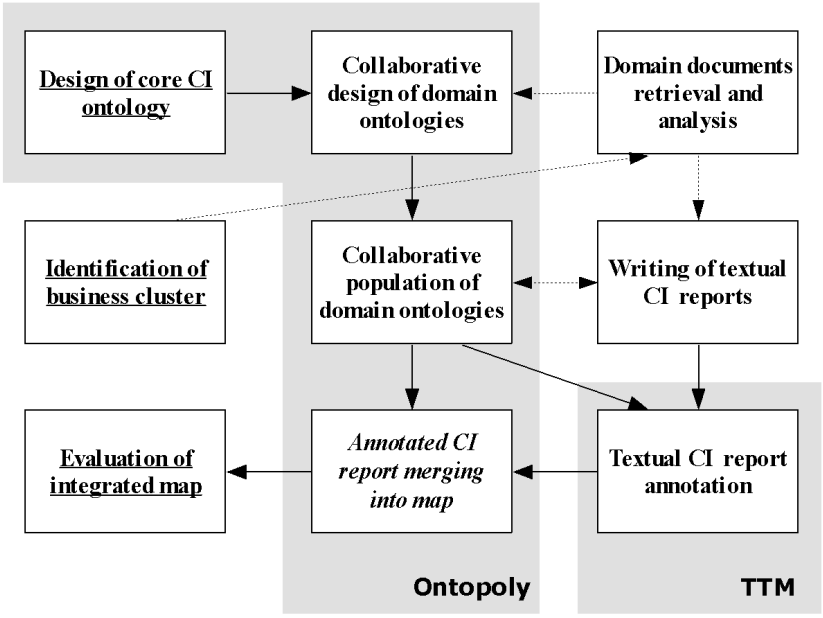

Figure 1: Schema of workflow

ing them can also be shared to a similar degree. This is particularly important for SMEs.

In the paper we discuss the interrelationship between CI (as specific branch/variation of BI) and semantic technology, and report on an ongoing project on 'semantization' of CI reports in the context of business clusters. Section 2 presents the rationale for and some generic problems of the use of semantic technologies for CI. Section 3 digests the core facts on our project. Section 4 discusses the underlying domain models. Section 5 explains the complex workflow of activities leading to the creation and exploitation of semantic CI reports, as we perceived it in our project. Section 6 refers to the major software tools used in this workflow. Section 7 summarises some lessons learned from the first two rounds of the workflow. Finally, section 8 summarises the contributions of the paper and drafts future work.

2. CI AND SEMANTIC SYSTEMS

CI is sometimes understood as a separate discipline closely interconnected with BI, and sometimes (especially in Nordic countries) as inherent part of the BI proper. By our experience, applying semantic technologies in CI is even more critical than in 'classical' BI, as

1. external information is more heterogeneous and structurally richer than internal information, and
2. decision makers have to understand the new information in their familiar context of knowledge and reasoning.

Enriching CI reports with *semantic structures* is thus a natural way to support easier *retrieval* of relevant textual information by the decision makers (among other, via accommodating to their existing mindsets) and for creating *business maps* as well as *semantic portals* on the top of these maps. As CI reports are knowledge-rich but condensed documents, their 'semantization' is feasible through *authoring-based manual annotation*, though assistance by automated procedures is desirable. As the most important (not necessarily linear) steps towards a semantic repository for CI reports on a given domain and/or business cluster we consider: (1) ontology design; (2) ontology population; (3) ontology-based text annotation; (4) interlinking.

Current inventory of semantic technology (be it based on OWL/RDF standards or on the Topic Maps standard, which we adopted in the current project) however mainly focuses on processing elementary information on a detailed level of resolution. It is thus desirable to combine them with *professional analytic tools*, such as Analyst's Notebook[2], which are capable to detect complex relationships in structured as well as unstructured information.

3. PROJECT OVERVIEW

Within the joint effort of Tovek, as an SME specialised in knowledge technology and member of The Society of Competitive Intelligence Professionals (SCIP),[3] and the University of Economics, Prague (UEP), in the course of one academic year (2007-8), undergraduate students were trained to collect and assemble information relevant for CI goals as well as to master several knowledge technology tools.

A base of over 70 annotated CI reports arose by the coordinated effort of student teams; nearly 300 students got involved overall in the (joint) role of report writers, annotators and 'ontologists'. The average size of a textual report was about 3 000 words; there were, on average, several tens of annotations per report, each typically spanning over one or

[2]Product of i2 Ltd, see `http://www.i2.co.uk/products/analysts_notebook/`.
[3]`http://www.scip.org/`

Figure 2: Industry Ontology Excerpt

few sentences or paragraphs. Three domains, in which business clusters explicitly exist or can potentially be formed, were addressed: *packaging industry*, *glass industry* and *information industry*. Every cluster was examined from the point of view of about 20 key organisations.

4. DOMAIN MODELS

For each domain a specific domain ontology was built, taking a *core CI ontology* as start-up. In the first run of the experiment, each student team expanded the core ontology separately so as to accommodate the needs of their annotation activity. However, since automated ontology mapping tools are not sufficiently reliable for reconciliating such separately extended ontologies ex post (and manual mapping would be very tedious), in the subsequent experiments the student teams designed (extended) the ontology collaboratively from the beginning.

The original core ontology was taken from Tovek and was designed to suit one company only. This design proved to be unsuitable for collaborative creation of the ontology and, therefore, a new core ontology was designed to suit the whole industry (see a part of the ontology in Fig 2). In the end of the process, the collaborative effort resulted in three industry-specific ontologies, which evolved from the original core ontology. Each of these ontologies contained about 100 concepts that described each of the industries.

The underlying *CI model* for all three domain-specific studies was that of *Porter's Five Forces*, which is a business methodology for qualitative evaluation of company's strategic position [3]. In accordance with this model, the reports primarily focused on the following issues: the threat of *new entrants*, the bargaining power of *customers*, the threat of new *substitute products*, the bargaining power of *suppliers* and the rivalry of *existing competitors*.

5. SEMANTIC CI REPORT WORKFLOW

The workflow of semantic CI report creation, as it crystallized in the two iterations of the project, is depicted in Fig. 1: boxes correspond to activities, solid arrows to interdependencies involving direct data/artifact flow among activities and dashed arrows to interdependencies without direct data/artifact flow. The activities on the left-hand side (with underlined text) were carried out by CI experts from Tovek; the 'merging' activity in the middle bottom (with slanted text) was carried out by experienced knowledge engineers (and teachers) from UEP; all the remaining activities were carried out by UEP students under modest supervision of teachers. Two 'semantic' *software tools* were used: Ontopoly and Tovek Topic Mapper (TTM).

The initial impetus was from the CI experts who designed *core ontology of CI* (covering, in particular, numerous notions defined in Porter's Five Forces) and also suggested interesting *business clusters*. The student teams bid for companies from the given domain pool and then started to collect *relevant textual documents* such as news articles and web pages that were relevant with respect to 'their' company. Information collected from these resources was the basis for *writing textual CI reports*. At the same time, the students collaboratively extended the core CI ontology with *domain-specific concepts and relations* (see Fig. 2) and then *populated* it with *instances* such as companies, products or people and their interrelationships. The textual reports were then loaded into the TTM tool and manually *annotated* with ontology entities. A selected (by quality) subset of annotated reports was then *merged*, together with the underlying ontology, into a larger *CI map* allowing to access the full documents,[4] which was submitted back to the CI experts. The final phase, *evaluation* in the business context (among other, using tools such as Analyst's Notebook), is ongoing.

6. SOFTWARE SUPPORT

As mentioned above, two semantic technology tools are being exploited in the project: Ontopoly and Tovek Topic Mapper. Both tools use the lightweight semantic formalism of *Topic Maps*,[5] which was sufficient for our purposes and could be mapped to a more expressive formalism (such as OWL/RDF) in the future if needed.

Ontopoly[6] is a generic tool for editing and browsing Topic Maps ontologies; for collaborative ontology design and population it had to be, however, adapted so that students could remotely update ontology data stored on a PostgreSQL server.

Tovek Topic Mapper (TTM) is a freely-downloadable[7] tool for ontology-based text annotation developed by Tovek for the EU Calibrate project[8] and further adapted for the current project. Fig. 3 shows a screenshot of TTM. On the left-hand side is a CI report listing claims of a company. The respective section heading is annotated with the (more general) concept 'Assets' from the taxonomy displayed on

[4]E.g. by querying in Tolog, `http://www.ontopia.net/topicmaps/materials/tolog.html`.
[5]`http://www.topicmaps.org/`
[6]`http://www.ontopia.net/solutions/ontopoly.html`
[7]From `http://www.tovek.cz/produkty/topicmapper.html`
[8]`http://calibrate.eun.org`

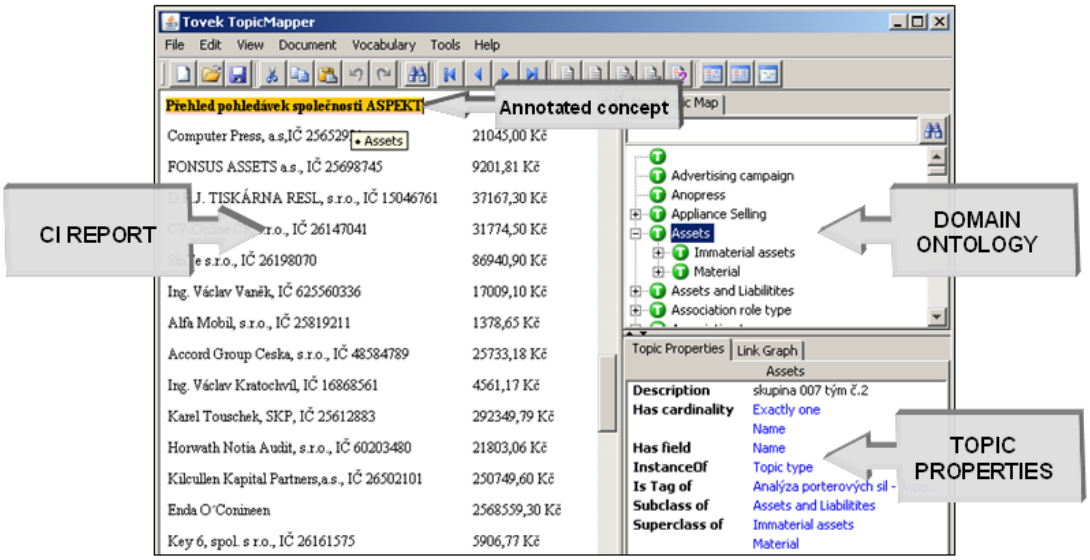

Figure 3: Annotation of a report in TTM

the right-hand side. TTM also supports other relations between concepts in the ontology. In principle, TTM allows not only to display but also to edit the ontology. However, in our workflow, the students were no longer permitted to edit the ontology in the phase of annotation by TTM (in order to avoid version deviation).

7. LESSONS LEARNT

The biggest obstacle we initially encountered in our project was the posterior alignment of ontologies created by each of the student teams. The Ontopoly tool offers string-based and PSI-based[9] alignment. We found out that PSI-based alignment is reliable but requires that annotation guidelines are in place and followed. Additionally, it is only usable for a small number of entity types[10]. In our experience, the string-based alignment (exact match) resulted into a large number of omissions, leading to duplicate entities in the merged ontology. This was the main reason for considering the option of collaboratively editing and maintaining the ontology.

We in fact adopted the *ontology maturing* paradigm [1, 2], and viewed ontology building as a maturing process that requires collaborative editing support and the integration into the daily work processes of the knowledge workers. The collaborative ontology design does not suffer from these shortcomings as there is only one copy of the ontology at a time. Each team contributed with the entities needed for the annotation of its report, provided such concept was not already present in the ontology. In this way, the ontology design reflected their work process. The case when the same concept would be added simultaneously by two students did occur, but only rarely. The resulting ontology nevertheless contained a few duplicates, which can be attributed to the failure to find an already existing concept or to different interpretation of the class hierarchy by the student. In our opinion this problem could be almost eradicated if the ontology design software 1) featured fast string-based concept search and 2) alerted the annotator upon insertion of a new concept if a concept with similar string representation already existed. We are now looking for alternative collaborative ontology editing tools for this purpose; an example is SOBOLEO, which is, unlike Ontopoly (as classical Web 1.0 application), a Web 2.0 tool utilizing AJAX technologies.

Another area of problems is the creation and annotation of the competitive intelligence reports. The students so far had free hands in choosing the text editor (most of them used MS Word), but they had to bear in mind that the result should be converted to HTML. The HTML file (which, when exported from MS Word contained lot of interfering HTML code) was then imported into Tovek Topic Mapper and transformed into the Topic Maps XTM format. The conversion however left the messy code unaltered and allowed it to enter the merged topic map. The bad readability of the HTML code is an obstacle to effective browsing and search in the collection of annotated files, for the target users. This issue is, however, now being subject of software tuning within Tovek, and will probably not affect the next runs.

8. CONCLUSIONS AND FUTURE WORK

The presented project is probably one of the first attempts to systematically apply semantic technologies in connection with textual CI report authoring, especially in the context of large business clusters. The ultimate goal of the project is to develop a methodology for efficient mapping of information

[9]PSI, Public Subject Indicator, is a subject indicator that is published and maintained at an advertised address for the purpose of facilitating topic map interchange and mergeability.

[10]In our case, these were only Czech companies: these can be unambiguously described by a PSI containing their national id number.

about the competitive environment aiming at:

- fast *retrieval* of relevant information in order to support operational decisions, as well as
- lucid *presentation* of complex situations in order to support strategic decisions.

As a technological side effect, the project may also serve as generic testbed for collaborative ontology design; this nowadays popular approach[11] has probably not been extensively tested in connection with the Topic Maps formalism yet.

An inherent problem of the study is the reserved attitude of some members of business clusters to joint CI undertakings (be they based on public resources) in general and to the use of semantic technologies for this purpose in particular, which makes the industrial feedback rather lengthy.

On the other hand, there is ample room for improving the quality of results, which would presumably lead to lowering the barriers between the academic project and the business clusters. Several updates will be effectuated in the next round: the quality of ontology design and population should rise thanks to more substantial training of students in *ontological engineering*; the form of annotations will be more uniform thanks to the availability of *annotation guidelines* (dealing with granularity issues etc.); a *content management system* will help manage documents more easily; finally, a *named entity recognition tool* will assist the students, allowing to create annotations more rapidly.

Furthermore, there is an ongoing work on incorporating the collaborative ontology design feature to the TTM annotation tool. This will allow to include new concepts to the ontology *on the fly* during the annotation, in the situations when the ontology does not contain the desired concept. Since there is one common ontology repository now, this will not reproduce the problems with multiple versions of the same ontology we experienced in the first iteration of the experiment.

Last but not least, the presentation of results is a critical issue. Based on our further work with Ontopia Knowledge Suite[12] we can now publish the results as a Web 2.0 enabled semantic portal using the Topic Maps technology as a background; this would ease the delivery from the academic research labs to the decision makers' tables.

We believe the semantic workflow, which is the output of our project, in combination with the Tovek Topic Mapper and the Ontopia Knowedge Suite is in the stage in which it is applicable as a case study for education purposes both on undergraduate- and graduate-level knowledge engineering courses. Clearly, the required amount of manpower prevents this from large-scale adoption in business environments. Nevertheless, the technology and the workflow is available and could be used in mission-critical applications, where budget is not the main constraint.

[11] Cf. `http://km.aifb.uni-karlsruhe.de/ws/ckc2007/challenge.html`

[12] `http://www.ontopia.net/solutions/products.html`

9. ACKNOWLEDGMENTS

The research was partially supported by the the CSF project no. 201/08/0802 and partly by grant IGA 21/08 of UEP, Prague.

10. ADDITIONAL AUTHORS

Additional authors: Jan Rauch (University of Economics, Prague, email: `rauch@vse.cz`) and Marek Nekvasil (University of Economics, Prague, email: `nekvasim@vse.cz`).

11. REFERENCES

[1] S. Braun, A. Schmidt, A. Walter, G. Nagypal, and V. Zacharias. Ontology maturing: a collaborative web 2.0 approach to ontology engineering. In *Proceedings of the Workshop on Social and Collaborative Construction of Structured Knowledge at the 16th International World Wide Web Conference (WWW 07), Banff, Canada*, 2007.

[2] S. Braun, A. Schmidt, and V. Zacharias. Ontology maturing with lightweight collaborative ontology editing tools. In N. Gronau, editor, *4th Conference on Professional Knowledge Management - Experiences and Visions, Workshop on Productive Knowledge Work (ProKW 07)*, volume 2, pages 217–226, Berlin, March 2007. GITO.

[3] M. Porter. *Competitive Advantage*. The Free Press, New York, 1998.