# Building and Integrating Competitive Intelligence Reports Using the Topic Map Technology

Vojtěch Svátek[1], Tomáš Kliegr[1], Jan Nemrava[1], Martin Ralbovský[1], Jiří Šplíchal[2], Tomáš Vejlupek[2], Vojtěch Roček[1], and Jan Rauch[1]

[1] University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
e-mail: {svatek,tomas.kliegr,nemrava,ralbovsm,rauch}@vse.cz
[2] Tovek spol. s r.o., Chrudimská 1418/2, Praha 3, Czech Republic
e-mail: {splichal,vejlupek}@tovek.cz

**Abstract.** Competitive intelligence 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.

## 1 Introduction

*Competitive intelligence* (CI) is an ethical business discipline that supports decision makers in understanding the competitive environment. Its main vehicle are *CI reports*, 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.[3] 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 exploiting them can also be shared to a similar degree.

We report on an ongoing project on 'semantization' of CI reports in the context of business clusters, by means of Topic Maps. An earlier phase of the project was presented in [1]; the contribution of the current paper is in introducing a content management system (CMS) into the CI report workflow, in the coverage of the last phase of the workflow, posterior evaluation of the integrated map, and in explicitly discussing issues related to Topic Maps modelling. Section 2 lists the core facts on the massive involvement of (business informatics) student teams, which is crucial for the project. Section 3 discusses the underlying models of problem domains as well as of CI. Section 4 explains the complex workflow of activities leading to the creation and exploitation of semantic CI reports, including some observations from the evaluation. Section 5 discusses issues related to the usage of Topic Maps as technology. Finally, section 6 summarises the contributions of the paper.

[3] http://en.wikipedia.org/wiki/Business_cluster

## 2 CI Efforts of Student Teams

Within the joint activity of Tovek, an SME specialised in knowledge technology, and University of Economics, Prague (UEP), in the course of three academic semesters (September 2007 – January 2009), 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 about 130 annotated CI reports arose by the coordinated effort of student teams; about 500 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. Four domains, in which business clusters explicitly exist or can potentially be formed, were addressed: *water management*, *packaging industry*, *glass industry* and *information industry*. Every cluster was examined from the point of view of about 20 key organisations.

## 3 Domain Models and CI Model

For each domain a specific domain ontology was built, taking a *core CI ontology* (see Fig 1) 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, posterior mapping of ontology versions then was quite tedious. Therefore in the subsequent experiments the student teams designed (extended) the ontology collaboratively from the beginning.

The underlying *CI model* for all four domain-specific studies was that of *Porter's Five Forces*, which is a business methodology for qualitative evaluation of company's strategic position [2]. 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*.

## 4 Semantic CI Report Workflow

The workflow of semantic CI report creation is depicted in Fig. 2: boxes correspond to activities, solid arrows to interdependencies involving direct data/artifact flow among activities, and dashed arrows to interdependencies without direct such flow. The activities on the left-hand side (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. Three *software tools*, visualised by means of shadowed background, were used in different phases of the workflow: the Topic Maps editor *Ontopoly*,[4] a CMS based on *Joomla!*,[5] and the CI-specific software *Analyst Notebook* (AN).[6]

[4] `http://www.ontopia.net/solutions/ontopoly.html`
[5] `http://www.joomla.org/`
[6] `http://www.i2.co.uk/products/analysts_notebook/`

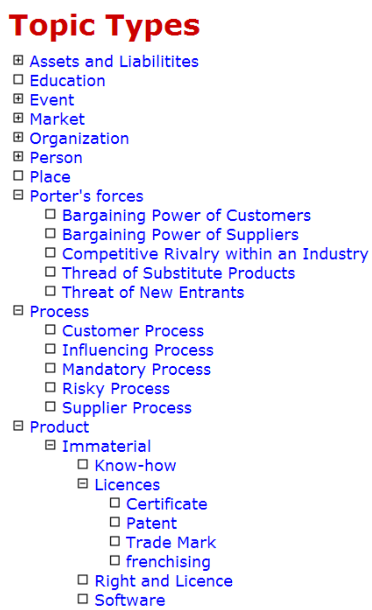


**Fig. 1.** Core CI Ontology Excerpt

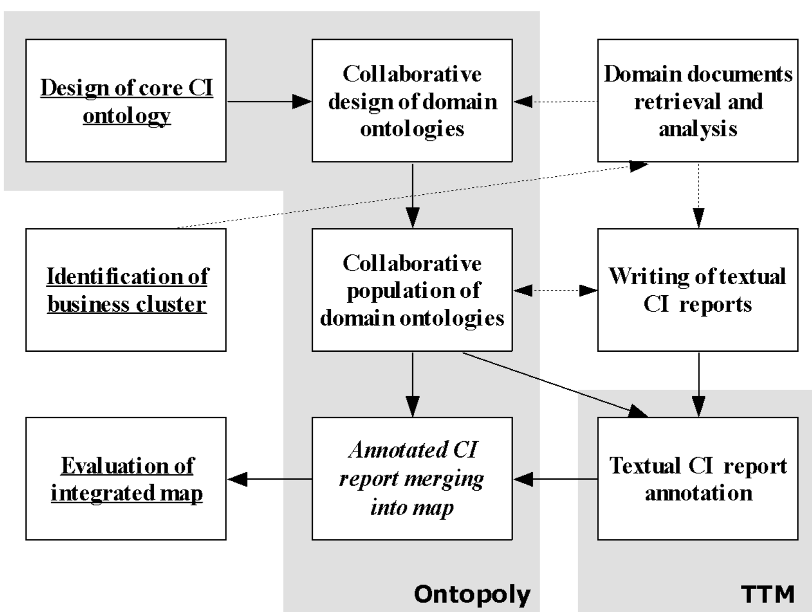


**Fig. 2.** Schema of workflow

The initial impetus was from the CI experts who designed the *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. 1) and then *populated* it with *instances* such as companies, products or people and their interrelationships. For the purpose of collaborative ontology editing and population, Ontopoly was adapted so that students could remotely update ontology data stored on a PostgreSQL server.

The textual reports created in MS Word, with automatically generated *tables of content* (TOC), were then loaded into the *CMS*; their fragments corresponding to TOC headings thus became addressable by URIs. These URIs were then stored in the topic map as *external occurrences*, the texts thus obtaining *fine-grained semantic annotation*. As a result, by clicking on the topic in a Topic Maps browser, the user can immediately navigate to the occurrences of the topic in the text and view the whole *context* of each occurrence. This approach (only used in the last semester) seems to be superior to the previous one described in [1], where the annotation was carried out entirely manually in a dedicated annotation tool and textual fragments were stored in internal occurrences of the topics. The new approach has many practical advantages to the old one: the text with full content (including even figures or links) in the CMS is more intelligible than fragments in internal occurrences, which were often a product of chaotic clicking; the CMS is better adapted than Ontopoly for storing non-Latin-1 characters (common in Czech); further editing of an article is possible in the CMS without invalidating the annotation; finally, the full-text search feature of the CMS can be exploited, while in the old approach the original Word documents were unlinked to the topic map and had to be searched manually.

The last phase, (to date, tentative) *evaluation* from the point of view of a final user of the CI map, was carried out by an advanced student with extensive knowledge of CI, in collaboration with experts from Tovek, for a single use case, namely, that of water management. The content of the integrated map was further enriched with information from a large, public business database, *Magnus*,[7] which characterises more than a half million organisations in the Czech Republic. The database is relatively poor in relationships but has large coverage of economic parameters of individual companies; this makes it an excellent complementary resource to the original topic-map-aware CI reports, which primarily contributed by relations among entities and by information about products/services. Eventually, the map was imported into the AN tool. A fragment of the map, showing interconnected companies (pentagonal shape), product categories (squares) and persons, as displayed by this tool, is in Fig. 3. It seems that business-oriented tools like AN, to which CI users are fully accustomed, can often be more suitable for detecting complex, high-level relationships in business data than generic tools that strictly adhere to a knowledge representation standard such as Topic Maps.

[7] http://www.magnus.cz/?idf=magnus-magnusweb

Experts generally found the integrated map useful for CI purposes. They however noticed the problem of *imbalanced density* of the map; while some companies were involved in dozens of links (to products, people, other companies, regulations etc.), some other were nearly isolated as nodes in the integrated map. Further effort will be needed to check how much of this phenomenon is due to inherent structure of the market and how much to uneven capacity of the students either in the report writing (and initial resource discovery) or semantic annotation phase.

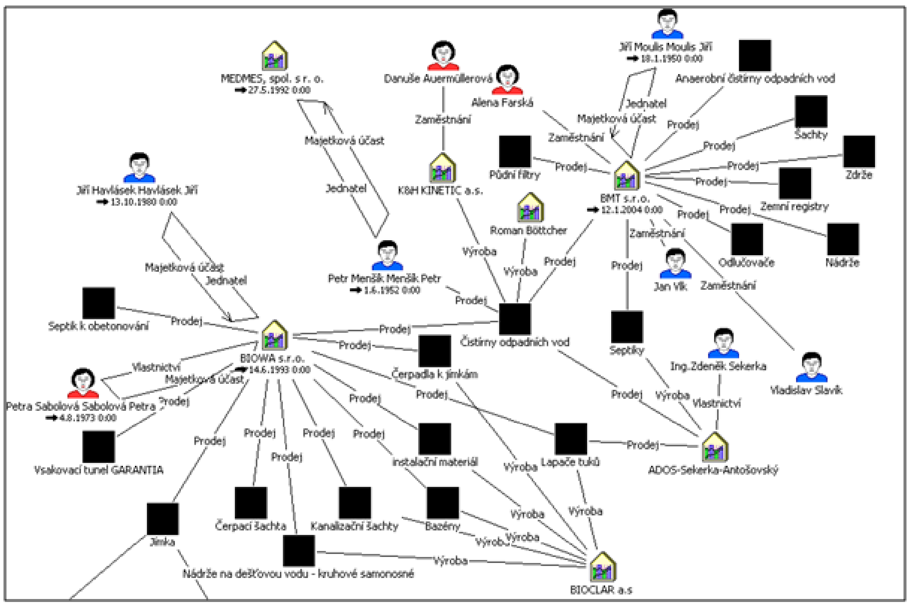


**Fig. 3.** Fragment of water management CI map in Analyst Notebook (labels are in Czech)

## 5 Topic Map Modelling Issues

The most critical caveat we initially encountered in our project was the *posterior alignment* of ontologies created by each of the student teams. Ontopoly offers string-based and PSI-based[8] alignment. 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.[9] In our experience, string-based alignment (exact match) resulted into numerous duplicate entities in the merged ontology. This was the main reason for shifting to collaborative editing of the ontology: there is only one copy of the ontology at

[8] PSI, Published 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.

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

a time, and each team contributes with the entities needed for the annotation of its report, provided such entity was not yet present in the ontology. The case when the same concept was added simultaneously by two students then occured much rarely. In our opinion this problem could be suppressed even more 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.

Although most students did not have prior experience with ontological modelling, they generally found Ontopoly easily usable and characterised the Topic Maps model as simple and intuitive. The most tricky modelling issues were the use of role types and relations with higher ($> 2$) arity. Students were typically unable to grasp the meaning of role typing and consequently left the role untyped. When modelling relations such as 'X competes with Y in product type Z', they usually reduced the complexity of the actual relation (e.g. to 'X competes with Y'). Having been reminded about the possibility of n-ary relation modelling, they found this notion useful. Unfortunately, only binary relations entered the final phases of the workflow anyway due to limitations of AN.

## 6 Conclusions

The presented project is 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 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 [3] has probably not been extensively tested in connection with the Topic Maps formalism yet.

We believe the semantic workflow, which is the output of our project, in combination with the Ontopia Knowedge Suite, a CMS such as Joomla! and professional analytical tools such as AN, is in the stage in which it is applicable at least as a case study for education purposes both on undergraduate- and graduate-level knowledge engineering courses. Moreover, even if the required amount of manpower prevents this approach from large-scale adoption in business environments, it could at least be used in mission-critical applications, where budget is not the main constraint.

*The research was partially supported by the the CSF project no. 201/08/0802.*

## References

1. Kliegr T. et al.: Semantic Annotation and Linking of Competitive Intelligence Reports for Business Clusters. In: OBI'08, Workshop at ISWC'08, Karlsruhe.
2. Porter, M.: *Competitive Advantage*. The Free Press, New York, 1998.
3. Tudorache T. et al.: Supporting collaborative ontology development in Protege. In: $7^{th}$ International Semantic Web Conference (ISWC-08), Karlsruhe.