Enterprise Information Integration (EII) with Topic Maps

Dipl.-Wirtsch.-Inf. Lutz Maicher

University of Leipzig, Augustusplatz 10-12, 04109 Leipzig

Abstract: Enterprise Information Integration (EII) is "the integration of data from multiple systems into a unified, consistent and accurate representation". One of the main requirements of EII is to get for each "thing of interest" one central information access point. But existing approaches fail caused by the absence of a semantic theory. Topic Maps are best suited for the management of large amounts of distributed information. And Topic Maps have a well applied theory about semantics: "One Topic for one Subject." My contribution is an (empirically tested) method for the creation of Topic Maps as semantic information integration layers in EII scenarios, based on the Topic Map theory.

Supervisor: Prof. G. Heyer, Abteilung für Automatische Sprachverarbeitung, Inst. für Informatik, Universität Leipzig, Augustusplatz 10, 04109 Leipzig

Deadline: November 2005

Research Gap – Semantic Interoperability in EII

Enterprise Information Integration (EII) "is the integration of data from multiple systems into a unified, consistent and accurate representation geared toward the viewing and manipulation of data" [Tayl04]. EII addresses one of the main business problems in information management: the disconnectedness of information and knowledge. Today, most information is available in digital form and *physically* connected by networks. However, in order to fully exploit the potential of digital connectivity, it has to be lifted to a *semantic* level. For example, only when all information regarding a special new technology, a project in progress, or a certain person is available at one central access point in an employee's working environment, connectivity has reached a sufficient level [Böhm+04].

The main approach to get such "seamless knowledge" [Pepp04] is to create views about these information sets where each "thing of interest" is represented by a central access point (proxy). A proxy provides access to all information about the "thing of interest". A set of all proxies has to fulfil the following requirements:

- **R1:** Act as a unified source of all available proxies in interest.
- **R2:** Provide a holistic view of all information regarding the proxies' "things of interest".
- **R3:** Declare the relationship between proxies and their "thing of interest".
- **R4:** Provide location transparency.
- **R5:** Enforce business rules to retrieve proxy segments in a given context.

As has been discussed in the literature [Pepp04], Topic Maps are well suited to fulfil all these requirements. Originally, Topic Maps are exchangeable back-of-book indexes. Today they are standardised by ISO with a focus on indexing any information in distributed heterogeneous environments.

A Topic Map is a set of Topics (proxies) and Associations (relationships between proxies). A Topic is the *central information access point* to all information of one Subject (the "thing of interest"). And a Subject is "anything […] regardless of whether it exists […] about anything whatsoever may be asserted […]. It is anything on which the creator of a topic map chooses to discourse." [TMDM]

The main semantic theory of Topic Maps is "One Topic for One Subject": inside a Topic Map only one Topic for one Subject is allowed. Whenever two Topics represent the same Subject they have to be merged to result in one single access point to all information: a seamless integration. This theory is well defined and applied within the Topic Map standards [TMDM, TMRM].

In comparison to the requirements listed above, a Topic Map is the *unified source (R1)* for all available information regarding all Subjects in interest. "One Topic for one Subject" assures that a Topic always provides a *holistic view (R2)* of a Subject. Because Topic Maps are overlaying information structures, *location transparency (R4)* is allowed. The concept of *Scope* allows the enforcement of business rules to *retrieve proxy segments in a given context (R5)* [Pepp01].

But "One Topics for One Subject" is only a theoretical (well defined and applied) framework which has to be put in practice carefully. Each Topic has to *declare a relationship to its Subject (R3)*. However, the common approaches solely operate on the syntactic surface (string identity). This is sufficient for most cases, but is insufficient for situations where Subjects aren't (commonly) named [Maic04]. How to handle Subjects which are abstract and have no sharp boundaries?

On an abstract level we encounter the same semantic problems like in natural languages: the relationship between *words* (Topics) and their *references* (Subjects). For exploiting Topic Maps' potentials, the problems of creating and using "languages" for the description of the relationship between Topic and Subject has to be discussed. If these issues of semantic interoperability are solved from the point of view of business information management, the use of Topic Maps will have huge opportunities in EII.

Research – Method for EII with Topic Maps

The aim of my research is the development of *a **method** for the creation of semantic information integration layers in EII scenarios based on the Topic Map theory*. Because of the gap between the semantic theory and its implementation discussed above, my main research is the development of a methodology for the creation and use of situation dependent "languages" which connect Subjects and Topics.

The method under development will meet Greiffenberg's requirements on methods as theories for the science of business information management [Grei03].

The theoretical basis of the method is a *Subject centric view* to EII. Important philosophical questions with respect to Subjects (and the semantics of the regarding Topics) arise (see [Kent78]): What is identifiable? What constitutes the boundaries of a thing in respect to its identity? Can identity evolve in time? Is identity situational or relative? How must properties of a thing change to alter its identity? How to name these Subjects? Or does a structuralist approach yield better results: the Subject of a Topic is primarily defined by the use of a Topic (its neighbourhood) inside a Topic Map? The answers to these questions will have to be integrated in the Topic Map theory.

The method under development will make explicit how Topic Maps should be produced in relation to information sets which typically occur in organisations: texts, process descriptions and relational data. The philosophical questions discussed above are thus put into the context of business information management scenarios.

How can the method empirically be tested? Created Topic Maps will be best suited for semantic information integration if they can be *merged* (or exchanged) easily. If Topic Map Fragments created by the method can significantly better be merged than others, the method is better suited for all kinds of integration tasks, even in distributed, heterogeneous scenarios. Hence, an advanced merging approach is being developed and used for an empirical evaluation [Maic04].

References

[Böhm] Böhm, Karsten; Maicher, Lutz et al.: "Moving Topic Maps to Mainstream - Integration of Topic Map Generation in the User's Working Environment." In: J.UCS (Springer), Volume 10, Special Issue I-Know 2004, pp. 241-251.

[Grei03] Greiffenberg, Steffen: „Methoden als Theorien der Wirtschaftsinformatik". In: Uhr, Esswein, Schoop (eds.): „Wirtschaftsinformatik 2003", Physica, Heidelberg, (2003), pp. 947-967.

[Kent78] Kent, William: „Data and reality". North-Holland, Amsterdam, (1978).

[Maic04] Maicher, Lutz: "Subject Identification in Topic Maps in Theory and Practice". In: Tolksdorf, Eckstein (eds.): "Berliner XML-Tage 2004", (2004).

[PEPP01] PEPPER, Steve: *Towards a General Theory of Scope.* Verfügbar unter: http://www.ontopia.net/topicmaps/materials/scope.html

[Pepp04] Pepper, Steve: Towards Seamless Knowledge - Integrating Public Sector Portals. In: Proceedings of Extreme Markup Language 2004.

[Tayl04] Taylor, John: „Thoughts from the Integration Consortium: Enterprise Information Integration: A New Definition", DM Review Online, (Sept., 2004).

[TMDM] ISO/IEC (ed.) : Topic Maps – Part 2: Data Model. Latest version available at: http://www.isotopicmaps.org/sam

[TMRM] ISO/IEC (ed.): Topic Maps – Reference Model. Latest version available at: http://www.jtc1sc34.org/repository/0544.htm