The Praxis of Social Knowledge Federation

Arnim Bleier, Patrick Jähnichen, Uta Schulze, and Lutz Maicher

Topic Maps Lab, Department of Natural Language Processing
Faculty of Mathematics and Computer Science, University of Leipzig,
Johannisgasse 26, 04109 Leipzig, Germany
{bleier, jaehnichen, uta.schulze, maicher}@informatik.uni-leipzig.de

Abstract. There are currently two streams that dominate the research on knowledge federation: The one is the trend towards Linked Data, leading to fine-grained structuring of information that is machine readable; The other is the reuse and co-creation of information that spreads the burden of its creation to the public and enables the availability of large knowledge corpora. In this contribution we outline the design principles and architecture of a prototype platform harnessing the praxis of user behavior to tackle issues of signal noise ratio as well as corrupting bits of information slashing the machine readability of such distributed generated contend.

Keywords: knowledge federation, Topic Maps, crowdsourcing, distributed content creation, virtual merging, internet sociology

1 Introduction

Most of the research on the future of knowledge management tends to assume, that a global leap in technological infrastructure is made. A particular example is the semantic web as envisioned by Berners-Lee in [1], which resonated strongly in academia but has failed to meet the expectations in practice, so far. This semantic web differs from the current version of the web by characteristics such as: strong separation of content and layout, structuring of information and shared vocabularies for markup as well as machine readability. The motive for the strong response from academia to the Semantic Web (presumably) is that the rolled out vision encourages hope for prototypes that unlock new applications and business models as well as pays off research costs in terms of return on invest. However that hasn't happened as anticipated, so far. One reason may be, that the assumption of a discrete and global technological leap is unrealistic. Another reason may be the unsolved problem of the signal noise ratio for machine readable content created by crowds. The goal of this workshop paper is twofold: First, to outline an approach towards a general praxis of knowledge federation that supports the creation of structured information, as well as its decentralized verification. Second, to apply this approach to a social knowledge federation platform that is currently under development by the authors.

2 Social knowledge federation

2.1 The Social Praxis

Knowledge federation as used in the context of this paper is best characterized by the first part of Karabeg's twofold definition [2] as verb and as a noun:

> As an activity, knowledge federation means joining together multiple individual knowledge artifacts under a single identity. This may take any form, ranging from a simple

> subject-centric organization of those artifacts by using a topic map or a dialog map [...], to creating a new artifact from the fragments of existing ones [...], to uniting the individual artifacts under a high-level view [...]

In the praxis of using our platform, individuals create subject centric topic maps and merge them (together with their own annotations) under a new merged topic map, a federated organization of information is then the result of an iterative social merging process.

The characteristic of social praxis is that actions on the level of the individual are the result as well as the shaping force behind emerging collective structures. [3] In terms of social knowledge creation, it can be observed, that an individual user creates information artifacts, that in the next step enable their collective reuse and refinement (e.g. creation of new information artifacts). Moreover the social structures knowledge artifacts are increasingly backed up and encoded by electronic web applications shaping new social spaces.

Guenther [4] identifies three characteristics determining these new social spaces: The first is content, referring to the information artifacts themselves (i.e. in topic maps and their constructs). The second is the code, incorporating the application architecture and enabling its usage. The third is metadata, referring to the information about the information such as usage and provenience.

2.2 The User

As we have discussed, social web applications enable the user to create, reuse and annotate content. This happens in communities, i.e. in overlapping, as well as distinct groups, that share an interest in a domain as well as goals and expectations on how to use the application [5]. This results in different levels of segmentation of the created information artifacts, ranging from commonly shared and agreed knowledge to highly specialized domains and differing points of view. While some of these communities may be interested in government spending, others may be interested in issues of global warming; and even within a group interested in global warming, different aspects of data may be seen. To support such a kind of knowledge creation and federation, three types of user engagement have to be enabled: the creation of content, the merging of multiple existing content resources to a new one and the annotation of existing content resources. As a result, the users should be enabled to individually and iteratively merge and annotate available information resources. However, to make such a high level federation in practical scenarios work, we first have to consider some aspects of community generated content.

2.3 The Problem of Relevance and Validity

In the traditional media landscape, professional authors and journalists created information resources, commented on them and connected them. The advent of electronic media empowered almost every user to also create and publish content. However, this development enables not only a long tail of highly differentiated niche domains, but also raises questions on how the awareness of crowds is (to be) directed. The resulting problem refers to the signal noise ratio. [6]

Consequently, in the case of community generated content, it is a requirement to empower the user to distinguish between (for him) relevant and irrelevant content. Let us pose a question first: Is the relevance problem of information not already solved by the linking mechanism of the web? The answer is no, since this web linking mechanism only gives answer on how much attention a particular resource gets, not to how useful it is. Moreover the shift from low structured and document centric content to semantic structured content requires an even more rigorous filtering criterion than the usefulness, since the information in one document becomes apparent in all documents it is federated with. As an example imagine a scenario in which a “malformed” map is frequently referenced. This malformation leads (for example) to the merging of all the topics in the maps it is merged into. Clearly no one would like to federate this map, yet it is highly linked. Consequently a mechanism is needed, that not only indicates the potential usefulness of a information artifact, but also prevents highly federated resources from being corrupted by a single document. Keeping the metadata representing the users’ assumptions on how useful and trustworthy an information artifact is and the later utilization of these individual decisions can be used to share the burden of the decision whether to trust a particular resource.

Consequently, we intent to apply crowdsourcing as a mean to harness the individual decisions of the users of a map in Maiana to extrapolate on the relevance and integrity of the information contained. This democratizes the awareness guiding of crowds - useful information bubbles up, the not so useful one bubbles down - and allows the detection of corrupted information artifacts as early as possible.

3 Maiana prototype

The Maiana platform is a back-end technology, enhanced by a white label web front-end supporting the usage practices described so far.

3.1 Architecture

The information artifacts in our platform are conformant to the ISO standard 13250 Topic Maps Data Model [7], short TMDM. The TMDM is defined by seven constructs and their respective merging rules. These constructs namely are: topic map, association, role, topic, occurrence, name and variant. The merging rules are defined by conditions of equality and result, if meet, in the merging of the constructs. However, two modes of merging have to be distinguished in the praxis of using Maiana: hard merging and virtual merging. The hard merging takes place within a particular map and ensures the internal consistency upon CRUD operations on the constructs of a single map and is carried out by the topic map engine itself, whereas the virtual merging always takes place between whole maps and is triggered by the user’s action of merging information located in distinct maps. The technique behind this type of merging is, that instead of merging the actual constructs virtual constructs are created that play the role of proxy objects. These proxy objects contain a list of references to the actual constructs in the virtually merged maps.

As an example imagine two topics located in separate maps, but both representing the city of Dubrovnik. In case of virtual merging these two maps a virtual topic proxy is created exhibiting the union of properties known about Dubrovnik, but internally handling only references to them. This technique has the benefit of keeping the original information unchanged while enabling the tracking of the provenience of the information in the virtual maps.

Maiana is a social Topic Maps platform. Every user may upload topic maps to the system, mark them as public and downloadable, thus sharing his or her information resources with others. As Fig. 2 depicts, topic maps in Maiana are not always created from a serialized format that a user may have on the filesystem. It is also possible to access serialized topic maps that are available anywhere on the internet. Additionally Wikipedia and Dbpedia resources are automatically identified and transformed into topic maps using the mappify webservice[1]. Users in Maiana are able to follow each other, i.e. see what changes others are doing on their information resources, and to explicitly watch certain topic maps, being informed whenever something changes in them.

[1] http://ws.mappify.org/ by Heuer, L.

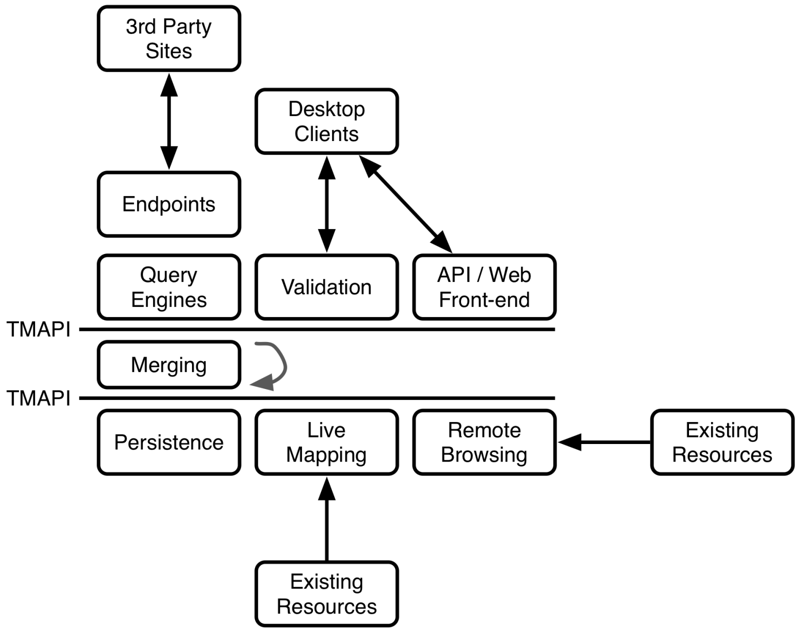

Fig. 1: **Layers and components of Maiana.**

While the web front-end of the platform is written in Ruby, the usage of JRuby allows the integration of highly performant java engines implementing the TMAPI. These engines handle tasks such as persistent storage and merging of topic maps, as well as live mapping of existing information resources to topic maps.

3.2 The Building Blocks of Maiana

- MaJorToM:
  The Merging Topic Maps Engine (MaJorToM) project has the goal to develop a lightweight, merging and flexible topic maps engine[2]. The engine provides the persistence layer for Maiana and exposes its functionality in an extended version of Topic Maps API [8], short TMAPI 2.0.
- Hatana:
  Hatana is the virtual merging engine enabling our platform to dynamically merge topic maps without actually editing any of them. The result of the merging process are virtual topic maps called containers. This dedicated merging engine works on the TMAPI 2.0 and exposes the TMAPI 2.0, resulting in a virtual Topic Map that is available for further reuse by components building on the TMAPI 2.0, including merging further merging. An additional useful characteristic of the technique of virtual merging is the track keeping of the provenience of information [10].
- TMQL4J:
  TMQL4J [9] implements the Topic Map Query Language, short TMQL [11], providing a

[2] http://code.google.com/p/majortom/

simple and straight forward way in Maiana to access and query for constructs and topic information organized in topic maps. The TMQL4J query engine is based on TMAPI and thus can also be used on a variety of topic map engines[3]. Besides query optimization and support for the current draft of TMQL, TMQL update queries are experimentally enabled and supported in the Maiana platform.

- Nikunau:
  Nikunau[4] is an implementation of the openRDF[5] Sesame Sail API enabling Linked Open Data Consumers to be integrated with the platform by the means of SPARQL and RDF/XML.[6]
- TMCL-Validator[7]:
  The Topic Maps Constraint Language (TMCL, ISO 19756) [12] is a language for the specification of constraints and schemas for topic maps. The TMCL validator validates a given topic map against a schema compatible to the current draft.
- (J)RTM[8]:
  JRuby Topic Maps [13] provides the glue between the different components in Maiana as well as the Ruby style language bindings used for the development of the Web Front-end.

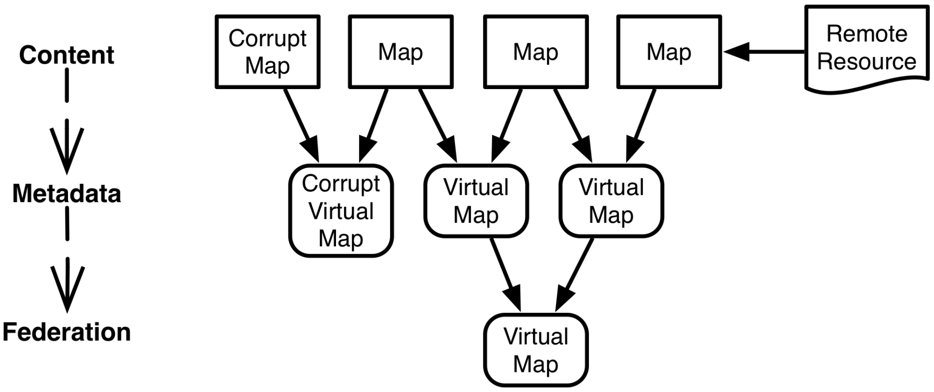

Fig. 2: **Merging of topic maps and remote resources into federated virtual maps.**

3.3 Ecosystem and Workflow

To give an example on how all these components can work together consider the following scenario: User A uploads two topic maps that contain information on John Lennon and George Harrison. Both maps are publicly available and are virtually merged into the virtual map C. As user B follows user A, it has come to his or her attention, that user A collects information about the band The Beatles. To collaborate on this, user B decides to create two other topic maps,

[3] http://code.google.com/p/tmql/
[4] http://bit.ly/aAKpsu
[5] http://www.openrdf.org/
[6] An up-to-date description of the Maiana SPARQL service http://is.gd/eTXF9
[7] http://code.google.com/p/tmcl-validator/
[8] http://docs.topicmapslab.de/rtm/

that take information about Ringo Starr and Paul McCartney out of Wikipedia or Dbpedia and merges these two remote resources in another virtual map D that also is publicly available. Now, user A just has to create a third virtual E in which the maps C and D will be merged and thus in which information about the band The Beatles is held. Because of the fact that all this information is merged only virtually, all changes that might be made to any of the underlying topic maps is directly reflected in the child and all its descendant maps. Moreover the information whether a certain piece of data originates from Wikipedia, Dbpedia or the two originally uploaded maps still persists and can is available. Since a virtual map is just topic map a it behaves exactly like any other topic map, making it possible to query it using TMQL or SPARQL as it would be done with any other information resource in Maiana.

To offer other developers the benefit we draw from Maiana, a public API has been released and implemented. Using this API, every registered user in Maiana also has access to his or her data from outside. This, besides the general behavior of topic maps to integrate with other data easily, makes it possible to use resources available in Maiana in one's own web or desktop applications[9].

Additionally some desktop applications that integrate with the Maiana platform have already been developed. For example Onotoa[10] offers a GUI to define the schema of topic maps via TMCL, whereas Genny [14] is a lightweight tool to create an editor for content creation given a schema.

4 Conclusion and Discussion

In this workshop paper we addressed the requirements and implementation of a prototype platform for social knowledge federation.[11] The contribution of our approach is the combination of subject centric Topic Maps with the social practice of gathering and harnessing metadata on their federation. On the Topic Maps side, Maiana acts as a stable repository for storage and reuse of the information in it, whereas on the social practice side, using Maiana adds value in terms of enabling crowdsourcing to merge meaningful maps but sort out malicious ones as early as possible. Only the synthesis of both properties enables the creation of highly federated views on content created by crowds.

[9] The current API description may be found at: http://bit.ly/bPVnm9
While the programming interface is continuously extended, new actions that explicitly deal with virtual maps will be introduced in the next releases.

[10] http://onotoa.topicmapslab.de/

[11] The current version of the Maiana platform is available at http://maiana.topicmapslab.de/

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American (2001)
2. Karabeg, D.: Knowledge federation patterns. In: Karabeg, D., Park, J. (eds.) First International Workshop on Knowledge Federation. vol. 552. CEUR-WS (2008)
3. Luhmann, N.: The Autopoiesis of Social Systems. In: John Benjamins Publishing ADVANCES IN ORGANIZATION STUDIES. vol. 14. pages 64–84 (2005)
4. Guenther, T., Schmidt, J.: Wissenstypen im "Web 2.0" – eine wissenssoziologische Deutung von Prodnutzung im Internet. Weltweite Welten (Jan 2008)
5. Schmidt, J.: Weblogs: eine kommunikationssoziologische Studie. Konstanz: UVK (2006)
6. Mell, J..: Signal vs noise. (2008)
   `http://jonmell.co.uk/signal-vs-noise/`
7. ISO/IEC IS 13250-2:2006: Information Technology - Document Description and Processing Languages - Topic Maps - Data Model. International Organization for Standardization, Geneva, Switzerland, `http://www.isotopicmaps.org/sam/sam-model/`
8. Heuer, L., Schmidt, J.: Tmapi 2.0. In: Maicher, L., Garshol, L.M. (eds.) Subject-centric Computing. Leipzig (2008)
9. Bock, B., Krosse, S., Maicher, L.: Topic maps run from xml and is coming back with flowers. In: Proceedings of XMLPrague (2010)
   `http://www.xmlprague.cz/2010/files/XMLPrague_2010_Proceedings.pdf`
10. Schulze, U.: Hatana - virtual topic map merging, Presented at: Sixth International Conference on Topic Maps Research and Applications. Leipzig (2010)
11. ISO/IEC WD 18048::2008: Topic Maps Query Language. International Organization for Standardization, Geneva, Switzerland,
    `http://www.isotopicmaps.org/tmql/tmql.html`
12. ISO/IEC IS 19756:2010: Information Technology - Document Description and Processing Languages - Topic Maps Constraint Language. International Organization for Standardization, Geneva, Switzerland, `http://www.isotopicmaps.org/tmcl/tmcl.html`
13. Bleier, A., Bock, B; Schulze, U., Maicher, L.: JRuby Topic Maps. In: Maicher, L., Garshol, L.M. (eds.) Linked Topic Maps. Leipzig (2009)
14. Niederhausen, N., Windisch, S., Maicher, L.: Generating an Ontology Specific Editor. In: Proceedings of the Fourth International Conference on Advances in Semantic Processing. Florence, Italy (2010)