A Corporate Tagging Framework as Integration Service for Knowledge Workers

Walter Christian Kammergruber
(Institut für Informatik Technische Universität München
Boltzmannstr. 3, 85748 Garching, Germany
walter.kammergruber@gmail.com)

Karsten Ehms
(Siemens AG, Corporate Technology
Otto-Hahn-Ring 6, 81739 München, Germany
karsten.ehms@siemens.com)

Abstract: Digitally supported knowledge work, using tags for content organization, creates inherent challenges. In this paper we show the design of a corporate tagging framework facing these challenges. We describe the implementation of a thesaurus approach as a lightweight alternative to a more sophisticated ontology design. An RDF based architecture with a Web 2.0 style editor enables average users to enrich social tagging data with semantic relations.
Key Words: tagging, orchestration, semantic, thesaurus
Category: H.5.4 D.2.10 H.3.5

1 Introduction

How do you keep your snippets, bits and pieces of information together? Today's knowledge work [Hube, 2005] is characterized by a multitude of larger information systems, smaller ICT tools and underlying file formats. Most creative — so called weakly structured — workflows stretch across systems and tools. Therefore tool supported knowledge work is often more kind of a hassle rather than an efficient flow [Csikszentmihalyi, 2002] of activities. With the advent of Web 2.0 tools in organizations (discussed as Enterprise 2.0 [McAfee, 2006]) at least granular hyperlinks and the capability to embed those into content, supports minimal integration allowing to switch from one application to another. In rare cases, the hyperlink can be complemented by dynamic linked information, e.g. through RSS or ATOM feeds. Still, cross application integration is far from being efficient. We refer to these problems as (personal) *orchestration challenge* [Ehms, 2010]. The term orchestration alludes to the requirement of composing and possibly configuring the tools needed for a certain task.

While this turns aforementioned workflows into "switch flows" between applications, challenges related to the organization of knowledge are not addressed by the mechanisms described so far. Typical Web 2.0 applications and more and

more (client sided) desktop tools inspired by the web provide *tagging* as the smallest common denominator for content organization. Tagging itself has some inherent shortcomings ([Kammergruber et al., 2010]) compared to more sophisticated "ontological" in vitro approaches. We call this the *semantic challenge* of social tagging. These drawbacks are multiplied by the orchestration challenge. Tags have to be re-entered and user assistance, such as auto completion, cannot benefit from tags stored in other systems. The same holds true for search, navigation and tag gardening [Weller and Peters, 2008] scenarios. The linkage between the semantic shortcomings of tagging and the orchestration challenge provides the rationale to tackle both in one approach delineated in section 3.

Main issues related to the orchestration challenge are: *(i)* A *growing number of tools* and systems being used in professional and private contexts, *(ii)* a *variety of technical storage formats*, partly proprietary, *(iii) different user interfaces* and underlying metaphors for interaction and finally, *(iv) heterogeneous ways of organizing information*, at least partly *not* linked to the semantic context of one application, but merely as a result of missing cross application metadata support.

Main shortcomings related to the semantic challenge are: Result sets to simple queries are incomplete because *synonyms* are not represented adequately. Ambiguous terms (*homonyms/ polysemy*) used as filters might deliver a huge amount of not relevant items. *Acronyms*, in general used as synonyms in a given context, help making domain related communication and information management more efficient. On the other hand there is an increased likelihood of polysemic clashes between terms. The latter is a problem when doing research in an open domain, again, leading to irrelevant query/filter results. Hierarchical or pseudo-hierarchical navigation, i.e. successive filtering or expansion, can only be provided if additional structural information (*hyponyms* or *hypernyms*) is present. These shortcomings can be described as a lack of explicit semantic relations between tags leading to bad precision or recall [Manning et al., 2008].

Of course it is not realistic, claiming to "solve" these mostly well known problems with yet another "magical" system. However we propose, that our tagging framework is an innovative approach, embedded into a "real world" corporate environment from the very beginning. Resolution of the sketched problems is at least possible for taggable systems, i.e. web based applications with permalinks and simple export mechanisms such as RSS or ATOM.

2 Related Work

What we call the *semantic challenge* is a problem addressed by a long list and history of scientific work. Especially in the area of research around artificial intelligence and in other fields such as information retrieval [Panyr, 1986] this

category of problems has received quite some attention. In our context we consider the concrete semantic challenge as the lack of structure between tags.

Braun et al. [Braun et al., 2007] describe an ontology maturing process, based on social tagging data, by which ontologies are created. In Soboleo [Zacharias and Braun, 2007] they provide a web based editing interface[1] for modelling concepts using SKOS[2]. Schmidt et al. [Schmidt et al., 2009] integrate this approach in a wider context related to personal and organizational learning.

A completely different approach is followed by projects such as MOAT [Passant and Laublet, 2008] or Faviki[3]. In these web applications the concepts of existing ontologies or similar structures are either mapped to tags or the concepts are directly used for tagging resources.

Weller et al. [Weller and Peters, 2008] follow with "tag gardening" a method for reorganizing folksonomies. Activities thereby include editing, re-engineering, manipulating and organizing tags. Social tagging data should become more productive and effective after the tag gardening procedure.

The *orchestration problem* is a central issue dealt with by personal information management tools such as Nepomuk [Groza et al., 2007] or Haystack [Karger and Jones, 2006], [Bernstein et al., 2007]. Nepomuk and Haystack are tools for data unification in personal information management. Their major target is interlinking pieces of information and thus making these pieces easier to retrieve when needed.

Having pieces of information scattered over different applications, is also discussed under the term "Unified PIM Support" [Jones and Bruce, 2005]. Lehel [Lehel, 2007] refers to users trying to solve those kind of problems as "information inventory control strategies".

3 Tagging Framework

In the following sections we first give a brief overview of the main characteristics of the architecture for the tagging framework. The last section focuses on the tag thesaurus editor being an essential part of the tagging framework.

3.1 General Technical Architecture

Figure 1 depicts a schema of the architecture. The tagging framework acts as mediator between different taggable applications (3). This addresses the orchestration challenge. The framework receives or fetches folksonomy data depending on the possibilities and implementation of the respective application to integrate.

[1] `http://tool.soboleo.com/editor/editor.jsp`
[2] `http://www.w3.org/2004/02/skos/`
[3] `http://www.faviki.com/`

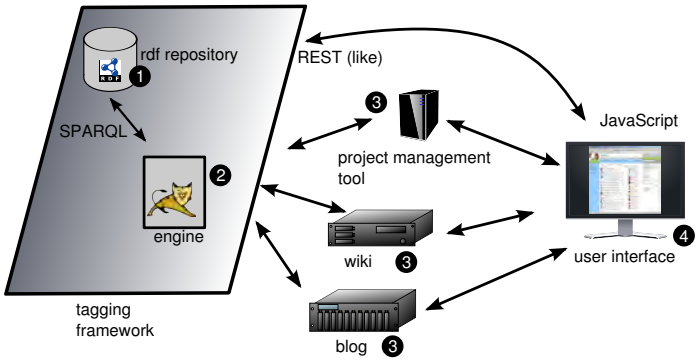

Figure 1: System Architecture.

The tagging framework consists of an RDF repository (1) and a servlet engine (2) acting as a container for the application logic. The internals of the tagging framework are transparent to the outside, i.e. the tagging framework can be accessed through a simple REST API [Fielding, 2000]. JSON[4] is used as default data serialization format. JSON can be, additionally to classical RPC mechanisms, natively processed with client side JavaScript (direct communication between (1) and (4)).

One major difficulty in the design of the architecture is the aggregation and especially the synchronisation of tagging data. We distinguish between two mechanism: "push" and "pull". Push means that a taggable application calls an API function from the tagging framework to inform the tagging framework that a change (create/ update/ delete) in its tagging data happened. "Pull" stands for a periodic fetch mechanism. The tagging framework triggers an update on it's tagging data for a certain taggable application.

3.2 Tag Thesaurus as Core Component

As described in section 1, folksonomies lack explicit formal structures. Therefore our goal is to extend a folksonomy with relations and to develop a thesaurus based on tags in an evolutionary manner. Figure 2 gives an overview of alternative vocabulary approaches (derived from [Weller, 2007]). The vocabulary types are ordered from left to right in increasing order depending on the potential depth of expressible semantic relations. Folksonomies are little more expressive than free keyword indexing since there is a social component included as well. For more details about the referred vocabulary approaches see [Gaus, 2005], [Peters, 2009], [Panyr, 2006].

[4] `http://www.json.org`

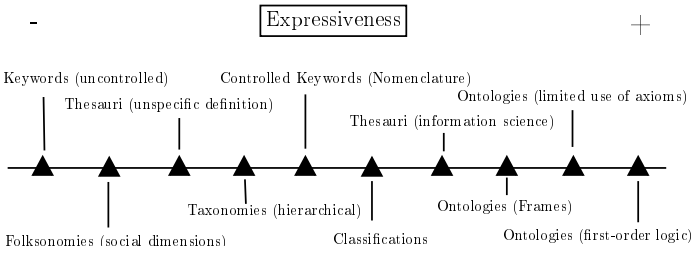

Figure 2: Expressiveness of vocabulary approaches (derived from [Weller, 2007]).

A thesaurus is a controlled vocabulary of terms that can be used as keywords. There are several variants of thesauri depending on the area they are used in. From a modeling perspective in general one can distinguish between two types of thesauri: "concept-oriented" and "term-oriented" ones. Concept-oriented means that entities in the thesaurus stand for an abstract meaning. Relations between concepts are expressed by links between concepts. Term-oriented means that term literals are interlinked directly.

There are several fields where thesauri find their application such as information science, biology or medicine. Sometimes these thesauri are a preliminary stage to an ontology and also referred to as one. The most widely used ones are linguistic thesauri since one is included in most popular word processors such as Microsoft Word or Open Office.

Recent work has proposed using social tagging data as a basis for an ontology [Braun et al., 2007]. We consider making a modest shift towards a term-oriented thesaurus being a more pragmatic solution. Since having too much complexity in the target model will most likely discourage an average user from participation. Furthermore we do not believe that a more complex model, such as a formal ontology, would provide enough additional benefit in navigation and filtering scenarios to justify the additional effort in modelling.

Out of the numerous possible thesaurus relations (see [Gaus, 2005] or thesaurus standard ISO 2788) we have selected four, which we believe are the most useful ones: *Synonym, Narrower, Broader* and *Related term*. Hence they are intuitively understandable by an average user and yet contain valuable semantic relations that can be exploited by our framework. In addition, we use a fifth relation (*Ignore*) by which a user can explicitly exclude any relation between two tags. These come in handy to overrule automatically proposed terms during query expansion and similar scenarios. Potential relations marked as *ignore* are excluded from further processing.

Figure 3 shows the user interface for the thesaurus editor. A user can define

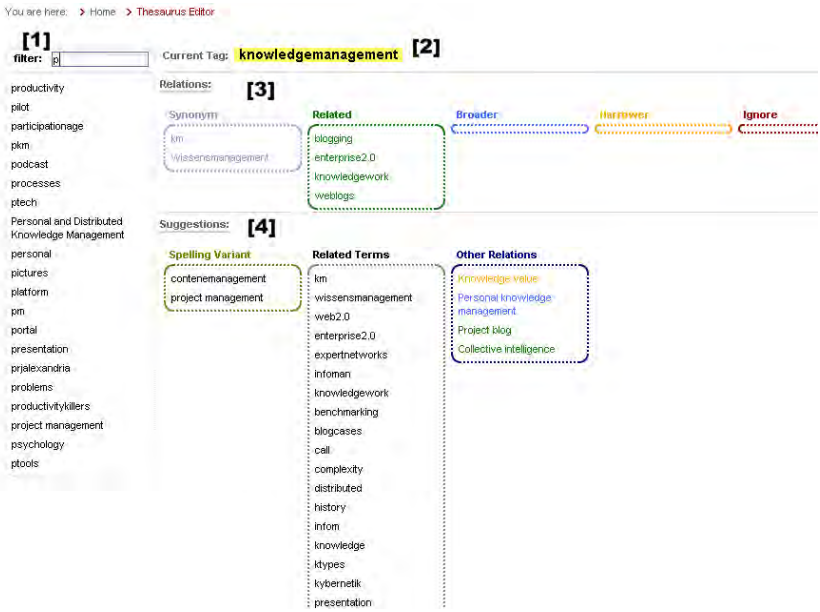

Figure 3: Drag and Drop thesaurus editor.

the semantic relations described above via drag and drop. One starts with selecting a tag from the folksonomy by applying a simple filter mechanism [1] (in this example the tag "knowledgemanagement" [2] is selected) which brings up already existing relations and related terms [3]. The bottom area of the screen [4] displays possibly related tags determined by different algorithms (string distance, tag co-occurrence, querying and mapping structured sources). The layout of the boxes suggests proximity between the results of certain algorithms and our thesaurus relations. The relations expressed by one user as well as user groups are stored in the RDF graph by a set of statements. The tag relation model specifies a multinary relation between a user having stated that two tags are associated by a certain *type* of relation (for details see [Kammergruber et al., 2010]).

4 Conclusion and Future Work

An instance of the tagging framework is currently under evaluation. We have tested a prototype in combination with several existing knowledge management services, such as global intranet applications (wikisphere [Lindner, 2008], blogosphere [Ehms, 2008]) and a project management tool.

Functional modules benefiting from the tag thesaurus and the tagging framework are amongst other: Tag autocompletion when searching for existing or cre-

ating new items, suggesting related tags, query refinement and expansion and tag clouds exhibiting relations between tags.

Having real world data allows us to assess the usefulness of the functionalities described. Our current experience with the framework in action has lead to plausible results. We are planning quantitative empirical analysis for instance validating relations between tags defined by users against corresponding Normalized Google Distances [Cilibrasi and Vitanyi, 2005]. Planned functional extensions include recommendations based on tagging data and/or social network analysis. First results have been published in [Kammergruber et al., 2009]

Acknowledgment

Parts of this paper are based on results of research within the framework of the Theseus project[5], more precisely the Use Case Alexandria. The project was funded by the German Federal Ministry of Economy and Technology under the promotional reference "01MQ07012".

References

[Bernstein et al., 2007] Bernstein, M., Kleek, M. V., Karger, D., and mc schraefel (2007). Information scraps: How and why information eludes our personal information management tools. *Transactions on Information Systems*.

[Braun et al., 2007] Braun, S., Schmidt, A., Walter, A., Nagypal, G., and Zacharias, V. (2007). Ontology maturing: a collaborative web 2.0 approach to ontology engineering. In Noy, N., Alani, H., Stumme, G., Mika, P., Sure, Y., and Vrandecic, D., editors, *Proceedings of the Workshop on Social and Collaborative Construction of Structured Knowledge (CKC 2007) at the 16th International World Wide Web Conference (WWW2007) Banff, Canada, May 8, 2007*, volume 273 of *CEUR Workshop Proceedings*.

[Cilibrasi and Vitanyi, 2005] Cilibrasi, R. and Vitanyi, P. M. B. (2005). Automatic meaning discovery using google.

[Csikszentmihalyi, 2002] Csikszentmihalyi, M. (2002). *Flow: Das Geheimnis des Glücks*. Klett-Cotta.

[Ehms, 2008] Ehms, K. (2008). *Globale Mitarbeiter-Weblogs bei der Siemens AG.*, pages 199–209. Oldenbourg, München.

[Ehms, 2010] Ehms, K. (2010). *Persönliche Weblogs in Organisationen – Spielzeug oder Werkzeug für ein zeitgemäßes Wissensmanagement?* PhD thesis, Universität Augsburg.

[Fielding, 2000] Fielding, R. T. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine.

[Gaus, 2005] Gaus, W. (2005). *Dokumentations- und Ordnungslehre: Theorie und Praxis des Information Retrieval*. Springer, Berlin, 5., überarb. a. edition.

[Groza et al., 2007] Groza, T., Handschuh, S., Moeller, K., Grimnes, G., Sauermann, L., Minack, E., Mesnage, C., Jazayeri, M., Reif, G., and Gudjonsdottir, R. (2007). The nepomuk project - on the way to the social semantic desktop. In Pellegrini, T. and Schaffert, S., editors, *Proceedings of I-Semantics' 07*, pages pp. 201–211. JUCS.

[5] `http://www.theseus-programm.de/home/default.aspx`

[Hube, 2005] Hube, G. (2005). *Beitrag zur Beschreibung und Analyse von Wissensarbeit.* PhD thesis, Universität Stuttgart.

[Jones and Bruce, 2005] Jones, W. and Bruce, H. (2005). A report on the nsf-sponsored workshop on personal information management. Technical report, The Information School, University of Washington, Seattle, WA.

[Kammergruber et al., 2010] Kammergruber, W. C., Brocco, M., Groh, G., and Langen, M. (2010). Collaborative Lightweight Ontologies in Open Innovation-Networks. In Hafkesbrink, J. and und Johann Schlichter, H. U. H., editors, *Competence Management for Open Innovation — Tools and IT-support to unlock the innovation potential beyond company boundaries*, pages –, Mühlheim an der Ruhr. to appear.

[Kammergruber et al., 2009] Kammergruber, W. C., Viermetz, M., and Ziegler, C.-N. (2009). Discovering communities of interest in a tagged on-line environment. In *CASoN2009: Proceedings of the 1st International Conference on Computational Aspects of Social Networks.*

[Karger and Jones, 2006] Karger, D. R. and Jones, W. (2006). Data unification in personal information management. *Commun. ACM*, 49(1):77–82.

[Lehel, 2007] Lehel, V. (2007). *User-Centered Social Software âĂŞ Model and Characteristics of a Software Family for Social Information Management.* PhD thesis, Technische Universität München.

[Lindner, 2008] Lindner, B. (2008). Der Einsatz von Wikis in der Siemens AG. I-KNOW.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schuetze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

[McAfee, 2006] McAfee, A. P. (2006). "Enterprise 2.0: The Dawn of Emergent Collaboration" . reprint 47306. *MIT Sloan Management Review*, 47(3):21–28.

[Panyr, 1986] Panyr, J. (1986). *Automatische Klassifikation und Information Retrieval.* Niemeyer Max Verlag GmbH.

[Panyr, 2006] Panyr, J. (2006). Thesauri, Semantische Netze, Frames, Topic Maps, Taxonomien, Ontologien – begriffliche Verwirrung oder konzeptionelle Vielfalt? *Information und Sprache. Festschrift für Harald H. Zimmermann*, pages 139–151.

[Passant and Laublet, 2008] Passant, A. and Laublet, P. (2008). Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China.*

[Peters, 2009] Peters, I. (2009). *Folksonomies. Indexing and Retrieval in Web 2.0 (Knowledge & Information: Studies in Information Science).* De Gruyter, 1 edition.

[Schmidt et al., 2009] Schmidt, A., Hinkelmann, K., Ley, T., Lindstaedt, S., Maier, R., and Riss, U. (2009). Conceptual foundations for a service-oriented knowledge and learning architecture: Supporting content, process and ontology maturing. In *Networked Knowledge - Networked Media*, volume 221 of *Studies in Computational Intelligence*, pages 79–94. Springer Berlin / Heidelberg.

[Weller, 2007] Weller, K. (2007). Folksonomies and Ontologies. Two New Players in Indexing and Knowledge Representation. In Jezzard, H., editor, *Applying Web 2.0. Innovation, Impact and Implementation*, pages 108–115.

[Weller and Peters, 2008] Weller, K. and Peters, I. (2008). Seeding, weeding, fertilizing. different tag gardening activities for folksonomy maintenance and enrichment. In Auer, S., Schaffert, S., and Pellegrini, T., editors, *Proceedings of I-Semantics'08, International Conference on Semantic Systems. Graz, Austria, September 3-5*, pages 10–117.

[Zacharias and Braun, 2007] Zacharias, V. and Braun, S. (2007). Soboleo - social bookmarking and lightweight ontology engineering. In *Workshop on Social and Collaborative Construction of Structured Knowledge (CKC), 16th International World Wide Web Conference (WWW 2007).*