ISWC 2008

The 7th International Semantic Web Conference

Katharina Siorpaes
Elena Simperl
Denny Vrandečić

Incentives for the Semantic Web (INSEMTIVE 2008)

October 26, 2008

The 7th International Semantic Web Conference
October 26 – 30, 2008
Congress Center, Karlsruhe, Germany

ISWC 2008

Platinum Sponsors

Ontoprise

Gold Sponsors

BBN
eyeworkers
Microsoft
NeOn
SAP Research
Vulcan

Silver Sponsors

ACTIVE
ADUNA
Saltlux
SUPER
X-Media
Yahoo

The 7th International Semantic Web Conference
October 26 – 30, 2008
Congress Center, Karlsruhe, Germany

ISWC 2008

Organizing Committee

General Chair
Tim Finin (University of Maryland, Baltimore County)

Local Chair
Rudi Studer (Universität Karlsruhe (TH), FZI Forschungszentrum Informatik)

Local Organizing Committee
Anne Eberhardt (Universität Karlsruhe)
Holger Lewen (Universität Karlsruhe)
York Sure (SAP Research Karlsruhe)

Program Chairs
Amit Sheth (Wright State University)
Steffen Staab (Universität Koblenz Landau)

Semantic Web in Use Chairs
Mike Dean (BBN)
Massimo Paolucci (DoCoMo Euro-labs)

Semantic Web Challenge Chairs
Jim Hendler (RPI, USA)
Peter Mika (Yahoo, ES)

Workshop chairs
Melliyal Annamalai (Oracle, USA)
Daniel Olmedilla (Leibniz Universität Hannover, DE)

Tutorial Chairs
Lalana Kagal (MIT)
David Martin (SRI)

Poster and Demos Chairs
Chris Bizer (Freie Universität Berlin)
Anupam Joshi (UMBC)

Doctoral Consortium Chairs
Diana Maynard (Sheffield)

Sponsor Chairs
John Domingue (The Open University)
Benjamin Grosof (Vulcan Inc.)

Metadata Chairs
Richard Cyganiak (DERI/Freie Universität Berlin)
Knud Möller (DERI)

Publicity Chair
Li Ding (RPI)

Proceedings Chair
Krishnaprasad Thirunarayan (Wright State University)

Fellowship Chair
Joel Sachs (UMBC)

INSEMTIVE

1st International Workshop on Incentives for the Semantic Web

http://km.aifb.uni-karlsruhe.de/ws/insemtive2008/

Introduction

> *"The original Scientific American article on the Semantic Web appeared in 2001. It described the evolution of a Web that consisted largely of documents for humans to read to one that included data and information for computers to manipulate. The Semantic Web is a Web of actionable information—information derived from data through a semantic theory for interpreting the symbols. The semantic theory provides an account of "meaning" in which the logical connection of terms establishes interoperability between systems. […] This simple idea, however, remains largely unrealized." (*Nigel Shadbolt, Wendy Hall, Tim Berners-Lee (2006). The Semantic Web Revisited. *IEEE Intelligent Systems*.)

One of the reasons for this state of affairs, almost seven years after the publication of the seminal article on the Semantic Web, has been always considered to be the lack of high quality semantic content. A critical mass of semantically annotated Web pages, semantically enhanced multimedia repositories, as well as business-relevant, widely-accepted ontologies would provide a feasible basis for the development of semantic applications of immediate added value for its users, and for the adoption of semantic technologies at industrial level. Despite a mature set of techniques, tools, and methods for authoring semantic content, one can observe very limited user involvement. The lack of semantic content and the missing engagement of users can be traced back to the missing incentive models incorporated by semantic technology. This is very contrary to the Web 2.0 movement which lives great popularity and a huge amount of user contributions. Even though, there are also many failing Web 2.0 tools, applications like Wikipedia, Del.icio.us, Flickr, YouTube, Facebook or LinkedIn generate enormous user interest and massive amounts of data. Each of those applications implements an incentive that motivates people to contribute their time and human intelligence.

The workshop addresses **incentives** for building the Semantic Web, i.e. achieving tasks, such as ontology construction, semantic annotation, and ontology alignment. It is intended as a networking event for discussing and brainstorming ideas for motivating people to contribute to semantic content creation. The workshop also seeks for original academic work in the respective field including:

- Motivations and incentives of several Web 2.0 applications
- Suggestions how those motivations can be applied in Semantic Web applications
- Incentive structures both within enterprise intranets and the open Web
- Games

- Tools exploiting collective intelligence and the “Wisdom of Crowds”
- Community-driven applications
- Monetary and non-monetary rewards
- Social applications in general
- Empirical studies on the usage of Web 2.0 or social Semantic Web applications

Relevance of the Addressed Topics

Semantic technology has by now reached a level of maturity where tools and methods allow performing many task relevant for the creation and usage of semantic content. However, the success is currently limited by a lack of semantic content and only limited user involvement. We can observe a sharp contrast between semantic content authoring tools and authoring tools in the Web 2.0 movement. This can be traced back to missing incentive models incorporated by semantic technology. The workshop addresses this gap and aims at instruments for fostering user engagement in semantic content creation and therefore boosting semantic technology.
Recently, several European research projects have been started that directly address the topics of this workshop, like ACTIVE, MATURE, or KIWI. Also, there are a number of industry enterprises forming around these ideas, like Metaweb, Twine, or True Knowledge. Therefore we consider now to be the right time for this workshop.

Organization Committee

Elena Simperl, STI, University of Innsbruck, Austria.
elena.simperl@sti2.at

Katharina Siorpaes, STI, University of Innsbruck, Austria.
katharina.siorpaes@sti2.at

Denny Vrandecic, AIFB, Universität Karlsruhe, Germany.
dvr@aifb.uni-karlsruhe.de

Program Committee

- Sinuhe Arroyo, University of Alcala de Henares, Spain
- Chris Bizer, Freie Universitaet Berlin, Germany
- Dan Brickley, FOAF Project, UK
- Peter Haase, AIFB, Germany
- Tom Heath, Talis, UK
- Nick Kings, BT, UK
- Eyal Oren, VU Amsterdam, The Netherlands
- Carlos Pedrinaci, Open University, UK
- Valentina Presutti, Institute of Cognitive Sciences and Technology (CNR), Italy
- Marta Sabou, Open University, UK

- Sebastian Schaffert, Salzburg Research, Austria
- Andreas Schmidt, FZI, Germany
- Hideaki Takeda, NII and University Tokyo, Japan
- Tania Tudorache, Stanford University, USA
- Ilya Zaihrayeu, University of Trento, Italy
- Valentin Zacharias, FZI, Germany

Towards a Constitution Based Game for Fostering Fluency in "Semantic Web Writing"

Chide Groenouwe, Jan Top

Vrije Universiteit Amsterdam {chide|jltop}@few.vu.nl

Abstract. The Semantic Web (SW) is still far from realising its full potential, partly because it is still lacking enough high quality SW representations of information. We argue that a step in the right direction is fostering people's capability to fluently create high quality SW representations of the information they generate during problem solving processes. To foster such a capability, we propose a game in which teams compete in creating the best translations of texts into SW representations. Although playing the game is in itself already a way to foster such a capability, we moreover pursue learning from the game which are the most successful translation strategies (embodied by "constitutions") so that they can also be used by people outside a game setting.

1 Introduction

The Semantic Web (SW), as envisioned by Berners-Lee, holds the promise of improving human collaboration, by increasing the transparency and reusability of representations of information, specifically in a computational sense [1]. The SW still may not be considered to be mature, in part due to a lack of the availability of high quality SW representations of information. We argue that the SW cannot come to full maturity without fostering the human capability of creating such representations, not only in specialists but also in information creators [2] [3].

This article focuses on fostering the capability of doing so *fluently* and *during the process of creating information for a specific purpose*. With fluency, or translating in real-time, we mean that thoughts are translated instantly, with minimum delay and duration, comparable with what many nowadays can approach with normal writing. An important scenario is people immediately sharing these representations of thoughts on the Web for others to reuse and extend, in this way approaching "collective thinking". An application area is improving scientific collaboration, within which much valuable information is not put to full use. Moreover, we assume that the capability of persons to translate their *own* information is beneficial because, among other things, there are not enough knowledge engineers to keep up with the rate at which information is produced.

For the purpose of fostering the mentioned capability we propose a *game* setting, because this stimulates people to participate and improve their strategies. In the basic form of the game, two or more competing teams translate the same text(s) into SW content in a fixed short time, after which teams challenge each others translations by posing questions that have a very specific answer based on the content of the text(s).

A question could for example be: give me the total number of inhabitants of all Asian countries mentioned in the text (imagine a geography text that mentions 25 countries from different continents). Each team then tries to construct an algorithm that derives the answer from their own translation. The final score is based on the complexity of the algorithms and the correctness of the answers.

From the game we hope to learn which strategies and conditions (in the game embodied by "constitutions") are most effective for transforming thoughts (embodied by the translation process) as quickly as possible into high quality SW representations (embodied by representations with high scores). Moreover, we believe that the game will provide strong direct and indirect incentives for people to participate in the construction of the SW, including: (1) Playing the game itself makes people create valuable SW content instantly. (2) Reaching fluency (implying low cognitive strain) and experiencing the benefits of high quality SW representations during the game is a great incentive to also apply the acquired capability outside the game. (3) Successful constitutions can be used by others – also outside the game – as an example, which makes it easier to also acquire the capability and apply it.

1.1 Related Work

Games with a purpose were introduced by von Ahn to seduce people to volunteer enriching the Web with new representations of information that cannot be created automatically, by wrapping this purpose in appealing online games [4]. Von Ahn metaphorically speaks of "human computation", and putting lost "human computer cycles" into use. In the *OntoGames* project, Siorpaes and Hepp have adopted the same approach for creating SW content, and achieved promising results [2]. A difference with our work is that the focus of OntoGames is mass participation, with the disadvantage that the game must not be too difficult to play and some SW content authoring tasks must be sacrificed to make the game attractive for many. We assume that this limits the games to the creation of fairly "lightweight" SW content. However, creating deeper SW content is also essential to the quality of the SW, and this is what we focus on. If you want to put more of the "lost cycles" of the collective human computer to use, we argue that it is best if the cycles on the human computers with certain gifts would be spend on playing (much) tougher games with higher benefits, even if such human computers would form a minority. Moreover, a capability that seems not likely to be acquired by a majority *at this stage*, can still become so in the future. Note that conventional literacy grew explosively in less than two centuries, for example in a modern Western country as France from around 30% of its population around 1800 to above 95% in 1910, amongst other things as a consequence of improved methods of dissemination of the art of writing [5,6]. Our games could contribute to accelerating such a process for "SW literacy".

In the area of Computer Supported Collaborative Work we see a strong relation with work of Buckingham Shum et al: *Compendium* [7] and related systems such as *Claimaker* [8]. Buckingham shum et al are also promoting a form of digital literacy during collective sensemaking which has many similarities with ours [3,9]. However, their literacy is different in that it is primarily used as a means for a community to grow in their understanding of the topics they are dealing with, instead of increasing algorithmic transparency.

1.2 Overview

In 2 we develop preliminary notions and explain how this work extends our previous work with Open Constitution Based Knowledge Communities. In 3 the game designs will be explained, including the motives behind the design decisions. The work is concluded in 4.

2 Preliminary Notions and Previous Work

Before presenting the game designs in the following section, we will first explain how it extends our previous work in 2.1 and sketch preliminary notions on symbolic representation and reasoning in 2.2.

2.1 Open Constitution Based Knowledge Communities and Experiments

The overarching research project of which the games are part is the OCBKC-approach (Open Constitution Based Knowledge Communities), applied to fostering the mentioned capability [10]. In OCBKCs the way of collaborating, including a specification of the technology used, is written down as explicitly as possible in a *constitution*, which all participants of the community agreed on following. The advantages of a constitution based approach are: (1) The constitution integrates the technological and the human dimension in one whole instead of isolating both dimensions from each other. (2) The constitution can be used descriptively to allow an active reflection on the way of collaborating, and therefore a way of improving it. (3) Vice versa, it can be used normatively to experiment with certain ways of collaborating.

The constitution is moreover *open* in two senses: (1) Participants are stimulated to participate in the construction of their own constitution. We believe users participating in the construction of their own environment, instead of being imposed an environment, to be a crucial factor in fostering collective intelligence, as has also been suggested by others [11,12,13,14,15]. (2) The constitution is open for reuse by other people, so that they can benefit from it, adapt it to suit their own purposes (diversification) and participate in its improvement (evolution).

In our case, we wanted to create a constitution to foster the capability that is central to this article, and developed an initial version, which we subjected to experimentation with volunteers who agreed on collaboratively solving a simple problem, while following this constitution. This constitution also included the development of a complete software environment which we coined *Constitution Based Subleme*. Although the experiment turned out to be quite successful and the results promising, it appeared that translating the own thoughts, although being the final goal, had as drawback that it is difficult to steer the process towards certain modelling problems, and with this shape and refine certain parts of the pursued capability [16]. Therefore, we decided to offer participants texts to translate instead of translating their own thoughts. Moreover, we wanted to incorporate a game element, for the reasons mentioned before.

2.2 Symbolic Representation and Reasoning

First we will provide our view on some key notions concerning the algorithmic aspect of the SW, which are important for explaining some design decisions of the game. The system of interpretation of any SW Knowledge Base (KB) can be divided in two parts, an "algorithmic part" and a "purely human part". The algorithmic part defines which information extracting algorithms ("reasoners") are valid, even if these algorithms have not been written yet. The definition has been inherited from symbolic logic, which defines validity in terms of truth preserving transformations on the expression of the language [17] [18].The human part consists of the way humans (should) interpret the expressions in the KB.

The current common practice on the SW is as follows. W3C publishes a range of different partial systems of interpretation, each of which predominantly include an algorithmic part, currently RDF, RDFS, OWL Lite, OWL DL and OWL full. A person who wants to express information can then adopt one of these partial systems, for example OWL Lite, in his system of interpretation, and extend the human part as suits him/her (by introducing new vocabulary), as long as (s)he does not violate the semantic conditions of the integrated partial system and as long as (s)he doesn't extend the algorithmic part. An advantage of this practice is that it is much easier to reuse algorithms: the person in the given example can just apply all algorithms written for OWL Lite with confidence. A disadvantage is that people cannot locally extend or modify the algorithmic part of the system of interpretation so that more information can be extracted by means of algorithms.

3 Game Designs

This section presents two games variants: the full game and the simplified game. The first, described in 3.1 and motivated in 3.2, is the ideal setting, which, however, is difficult to realise on short-term. The second, explained in 3.3, will be a simplified version, which we intend to realise on short-term.

3.1 Full Game

The full form of the game is as follows. From a pool of people who are logged into the game, the computer randomly composes two teams, each of which consist of *Translators* and *Answerers* (they do not overlap). Moreover, it randomly chooses a text from some large database of texts. The Translators first agree on a constitution to use for the translation (see 2.1). They then translate the text into a digitised form they think is as optimal as possible for applying algorithms to it to answer any question that can be answered *precisely* and *unambiguously* solely based on the information contained in the text. There is a maximum time available for making the translation which is set as short as possible, forcing the participants towards fluency. Each team (both Answerers and Translators) then challenges the other team with posing a fixed number of questions. Subsequently, the Answerers get access to the final translation of their team and the list of all questions. For each question they try to develop an algorithm, which, when

applied to the translation, extracts the answer from it. The final score is based on the quality of the answer and the simplicity of the algorithm.

Details about the game are as follows: (1) The text is divided into fragments that do not overlap and each fragment is assigned to exactly one Translator. (2) The Translators produce two things: a translation and a definition of (the system of interpretation of) the language they used (see 2.2). The language definition may not express any information that is expressed in the text. For example, it could be a combination of RDF, RDFS, OWL DL and a set of own language extensions, including extension of the algorithmic part as defined in 2.2 . (3) The quality of the algorithm is based on its length after some normalisation (among other things, counting labels as one sign). (4) The algorithms that are produced by the Answerers, may not contain the information needed to answer the question. The simplest way to cheat would for example be writing an algorithm which contains the answer in the form of a string. This is indeed an algorithm that yields the answer, but the information is contained in the algorithm and not extracted from the information base. All teams have to publish their algorithms so that they can be scrutinised by the other teams and anyone else. (5) The constitution may be one that already exists (defined by a previous team), a modification of such a constitution, or an own constitution. The constitution is open, for any other person to reuse or scrutinise. The latter prevents cheating, by for example using the rules to communicate. If the constitution is new, it is added to the pool of constitutions from which future players may choose. The constitutions also will be ranked, based on the success teams using it had with it.

3.2 Motivation, Strengths and Weaknesses

Motivations for the game design decisions include the following. Assigning a text fragment to a single person approaches the situation of a person working with a local purpose, for example designing an experiment. The questions being about the text *as a whole* will create the necessity for the team of Translators, in spite of their local purpose, to put everything to work to make the *aggregate* of their local individual translations an algorithmically transparent whole. The questions being invented to attack the other teams, creates a strong incentive to make the questions as difficult as possible, and thus, the test of quality as good as possible. Composing the teams randomly creates the necessity for the team to at least make their way of collaborating explicit in the constitution, because they have no other way to coordinate their collaboration. (If they would know each other they could meet and train together and develop the coordination strategy off-line.) In this way, successful constitutions can be harvested from the game, to be reused by other people, in the game and for serious purposes. Allowing the Translators to define their own language, including the algorithmic part of the system of interpretation (and not being tied down by for example only being allowed to use RDF+RDFS) potentially allows all text that can be made algorithmically transparent to be made so.

Weaker points of the design include: how to cope with texts with internal inconsistencies (possible solution: not allowing them); there is only a necessity to make coordinative activities explicit in the constitution, not individual strategies; computational complexity is completely disregarded (possible solution: also incorporate exe-

cution time and memory consumption of the algorithms in score); bad Answerers can still ruin the score even when the produced KB is of high quality (possible solutions: (1) continue the competition with a randomly chosen set of new Answerers who get offered the translation and repeat this process a number of times, so that the translation can earn a long-term score, or (2) instead of randomly composing the teams, the computer looks at the average scores to combine people (for example combining Translators who played in teams with high scores with Answerers who played in teams with high scores, as to make strong teams, or (3) the game providers can build in tests with good translations and predefined questions with solutions stored in the system to identify the less skilled Answerers, an approach similar to one mentioned by Siorpaes and Hepp [2].

3.3 Simplified Game

We will now present the simplified version we intend to realise on short-term. It is equal to the full game except for the following modifications: (1) *Fixed algorithmic part for all: RDF + RDFS*: the part of the system of interpretation of the language that may be assumed to write algorithms is fixed to be RDF + RDFS for all competing teams. In future variants of the simplified versions we are considering other fixed parts, for example OWL DL. An advantage is that there is already quite some algorithmic support for this language, which makes it much easier to construct algorithms. For example, in case the algorithmic support chosen is a reasoning engine, the final algorithm could be equal to a query in combination with that reasoning engine. Disadvantage is that only a fraction of all information that could be answered with the help of algorithms can be answered with algorithms under this condition. (2) *Fixed constitutions:* instead of allowing teams to create their own constitution, we will offer them to choose from a fixed set of constitutions we developed. We will partly explain one of them in the following paragraph. Advantage: we do not expect that players will develop their own high quality constitution anytime soon, while ours is the result of substantial research and development, providing the players a head start. (3) *Preselected texts:* instead of random texts from a large database, we will choose specific texts that confront the teams with certain translation problems, amongst others: a text in which many different names are used for the same entity, so that the capability of reaching a shared vocabulary will be tested. Advantage: the number of participants will probably be relatively low in the time to come, and thus so the number of texts covered, and so the probability that the constitution is subjected to a range of crucial modelling problems when randomly choosing texts.

All fixed constitutions will be extensions of an adapted version of the constitution we used during our experiments in collaborative problem solving, as well the tool as the human part (see 2.1). We suffice with briefly describing a part of one of them due to limited space. The shared terminology is reached in the following way. When a Translator introduces a new node, (s)he must define it in natural language, according to certain criteria described in our earlier work [16]. These criteria include: being unambiguous; being generic; defining a single entity and not defining a node that has been defined earlier. The other Translators have to judge the definition as soon as possible. When someone doesn't agree with the definition of the node, it has to be revised, or the node

has to be withdrawn completely and other nodes must be proposed to express the given information. Only after reaching a shared status, the node may be used to express information (build triples). Moreover it contains rules concerning best modelling practices, such as how to cope with representing properties of sets of individuals.

4 Conclusion and Future Work

The main goal of this article was presenting a way to foster the human capability of creating high quality SW representations of information with great fluency during problem solving. For this purpose, we proposed a constitution based game, integrating a game element in our previous work with Open Constitution Based Knowledge Communities as explained in 2.1. The advantage of a game setting is that it stimulates participation and self-improvement. In the design of the game we wanted to incorporate products and lessons learnt from this previous work, among other things parts of the constitution (such as Constitution Based Subleme) and that it has and advantage to choose for text translation instead of translation of thoughts during problem solving, because the text offers a solid frame of reference to compare the results of different teams, and allows steering the process towards certain modelling difficulties. Two variants were presented: a full and a simplified game. Based on the argumentation developed in 3, we conclude that both variants are likely to help us approaching the main goal, the full game more so than the simplified.

As for our future work, we conclude that the simplified game is easier to realise on short-term. It will therefore be the one that we intend to implement first. Among the first participators in the game will be experimental researchers of Top Institute Food and Nutrition based in the Netherlands with whom close collaborations exist. Due to the high rate at which the experimental research community creates new information, it has much to gain from the pursued capability. We will report on lessons learnt from the execution of the games, in particular the elements that optimise the constitution, in subsequent publications.

ACKNOWLEDGEMENTS

This study has been funded by Top Institute Food and Nutrition (TIFN, http://www.tifn.nl/) and the Dutch Ministry of Economic Affairs (Ministerie van Economische Zaken, http://www.ez.nl/english/Organisation).

References

1. Berners-Lee, T.: Weaving The Web. (1999)
2. Siorpaes, K., Hepp, M.: Games with a purpose for the semantic web. IEEE Intelligent Systems (2008) `http://csdl.computer.org/dl/mags/co/2006/06/r6092.pdf`.
3. Buckingham Shum, S.: The roots of computer supported argument visualization. Visualizing Argumentation (2003)

4. Ahn, von, L.: Games with a purpose. Computer (2006) http://csdl.computer.org/dl/mags/co/2006/06/r6092.pdf.
5. Cipolla, C.: Literacy and Development in the West. (1969)
6. Vincent, D.: The Rise Of Mass Literacy. (2000)
7. Compendium: Visual hypertext concept mapping tool: http://www.CompendiumInstitute.org.
8. Buckingham, S.: Modelling naturalistic argumentation in research literatures: Representation and interaction design issues. International Journal of Intelligent Systems (22) (2007)
9. Selvin, A.: Fostering collective intelligence: Helping groups use visualized argumentation. Visualizing Argumentation (2003)
10. Groenouwe, C., Top, J.: Open constitution based knowledge communities in the semantic web. CEUR Workshop Proceedings **202** (2006) http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-202/.
11. Visser, J., Berg, D.: Learning without frontiers: Building integrated responses to diverse learning needs. (1999) http://www.learndev.org/dl/aect99-etrd.pdf.
12. Visser, J.: Learning communities: Wholeness and partness, autonomy and dependence in the learning ecology. (2001) http://www.learndev.org/dl/Barcelona2001.pdf.
13. Lévy, P.: Collective Intelligence: Mankind's Emerging World in Cyberspace. (1997)
14. Engelbart, D.: Augmenting human intellect: A conceptual framework. (1962)
15. Coolen, M.: De Machine Voorbij. (1992) To our knowledge only available in Dutch. Translation of title: Beyond the Machine.
16. Groenouwe, C., Top, J.: Real-time translation of thoughts into semantic networks during collective sensemaking: Quality of entity definitions. Technical Report V20080702 (2008)
17. Hayes, P.: Rdf semantics - w3c recommendation. (2004) http://www.w3.org/TR/2004/REC-rdf-mt-20040210/.
18. Dalen, v., D.: Logic and Structure. (1997)

Extracting Time and Location Concepts Related to Tags

Yukino BABA[1], Fuyuki ISHIKAWA[2], and Shinichi HONIDEN[1,2]

[1] The University of Tokyo
[2] National Institute of Informatics

Abstract. Folksonomy is a method of classifying content, and it is widely used in some web services. It allows users to choose tags (keywords or terms assigned to specific content) freely and to search content by referring to the tags. Compared to existing classification methods, folksonomy reflects the users' intention more directly because of its unlimited vocabulary and multiple tags for one content item. Moreover, it has a useful characteristic where tags represent the description of the content. Although tags are intended to be a rich semantic description of web content, machines cannot understand what the tags mean because they are just keywords. We describe a method to extract the concept related to the tag in a machine-understandable way by focusing on the features of content annotated with each tag. In particular, we target the problem of extracting the temporal and spatial concepts of the tags on Flickr, a popular photo sharing service by looking at the date and location distributions of photos for each tag. We evaluated the concept extracting method on a snapshot of actual Flickr data and show that it can identify a tags' concept in a manner similar to the way a person can.

1 Introduction

Folksonomy [1] is a method of classifying content, and it is widely used in web services (e.g., del.icio.us, Flickr, YouTube). It allows users to choose tags (keywords or terms assigned to a content) freely and to search content by referring to tags. Compared to existing classification methods, folksonomy reflects the users' intention more directly because of the unlimited vocabulary and multiple tags for one content item. Moreover, it has a useful characteristic where tags represent the description of content.

Meanwhile, the objective has been to build a Semantic Web where all web content contains machine-understandable metadata describing the meaning of each kind of content. However, the workload required for manual metadata creation and annotation is serious. Tagging could be part of the solution to this problem: A tag is utilized as a description of the content, but the tag is just a textual label, so machines are not able to understand the meaning. For example, when content is annotated with the tag Christmas, machines have no perception of what the tag means.

To solve this problem, we describe a method to extract the concept related to the tag in a machine-understandable way by focusing on the features of content annotated with each tag. For example, from the feature "Photos taken during December 24th and 25th are frequently annotated with the tag Christmas," we can get the information "The Christmas tag is related to a period during December 24th and 25th." In folksonomy, we can make a direct correlation between a tag and the features of content annotated with the tag because tags are added to content directly. Furthermore, we consider that the concepts obtained from tags better correspond to human recognition because of a better reflection of the people's intention, and we are likely to be able to extract the tags' concepts, which are difficult to determine using the top-down definition approach.

Using our method, for example, when content is annotated with Christmas and no date information is added, it is explicit that the content is related to a period during December 24th and 25th. This knowledge is useful to search content not only by keywords but also by time.

In particular, we target tags on Flickr [2], a popular photo sharing service that supports user-generated tags, and we extract tags with time and location concepts from the tags and information on the time and location at which the photos were taken. We define the time and location tags as being those related to time and location (e.g., time tags: August, cherry blossoms, Sapporo Snow Festival, location tags: Tokyo, Daibutsu, Sapporo Snow Festival[3]). Furthermore, we define the temporal and spatial concepts of time and location tags as being the ranges related to the tags (e.g., 2007-02-05 12:21:58 to 2007-02-16 23:48:24, (lat: 35.616279, lng: 139.650307) to (lat: 35.772702, lng: 139.859047)).

Our approach targets not only the tags that are explicitly related to the time or location (e.g., August, Tokyo) but also the tags that are implicitly related to the time or location (e.g., Cherry blossoms, Daibutsu). The latter tags' concepts are difficult to determine using the top-down approach, while the approach considering the features of tagged content as the users' intention is likely to extract the tags' concepts more easily.

We extract the temporal and spatial concepts by analyzing the distributions of annotated times and locations where the photos were taken. A method of determining whether the tags on Flickr are related to time and/or location has already been proposed [3]. However, this research did not extract the concepts of the tags. Our method extracts the concepts of the time and location tags.

The contributions are as follows:

- We provide an approach for extracting the concepts of time and location tags from the feature of the content annotated by the tags.
- We extend the existing method of determining whether the tags are related to time and/or location.
- We describe an application and analysis of this method for actual tags and photo information data taken from the snapshots posted on Flickr.

[3] Each tag can be both a time tag and location tag. Such tags can be considered as "event" tags, but we do not make exceptions for these tags in this paper.

We formalize our problem in Section 2. We describe the existing methods to determine time and location tags in Section 3 and describe a new method to extract the concepts of time and location tags in Section 4. Section 5 discusses the experiment, in which we evaluated our way of extracting concepts. We describe related work in Section 6 and conclude the paper in Section 7.

2 Problem Definition

In Flickr, each photo p has information on the date it was taken t_p and location it was taken l_p and some tags are annotated. We denote a tag generally as x and define certain classes according to the distribution of their temporal and spatial usage patterns as follows:

Time Tag The tag, the temporal distribution T_x of which has more than one dense region.

Location Tag The tag, the spatial distribution L_x of which has more than one dense region.

Moreover, we define concepts of these tags as follows:

The concept of the time tag The dense region in the tag's temporal distribution T_x.

The concept of the location tag The dense region in the tag's spatial distribution L_x.

The procedure of concept extraction is as follows:

1. Determine whether a tag is a time or a location tag; i.e., determine whether the tag has more than one dense region in the time or location usage distribution.
2. Extract the concept of this; i.e., the concept is taken to be the dense region if the tag's usage pattern has more than one dense region.

A standard pattern detection method, Naïve scan [4], perform steps 1 and 2. However, Scale-structure Identification [3], a method of determining whether the tags on Flickr are related to time and/or location, can perform step 1 with higher accuracy. Hence, in this paper, we present a way to apply Scale-structure Identification to step 2. In rest of this paper, we describe two methods to determine 1, Naïve Scan and Scale-structure Identification in Section 3. In Section 4, we show the methods to specify 2 with Naïve Scan and present the method to use Scale-structure Identification for 2, region specification.

3 Determining time and location tags

In this section, we describe two methods to determine whether a tag is a time tag or a location tag: Naïve scan and Scale-structure Identification. To simplify the discussion, we describe the methods only for time tags; the methods for location tags are similar.

3.1 Naïve Scan

Naïve scan is a standard algorithm in signal processing to detect a burst [4]. It divides the data's time distribution with scale value r and if the number of data in a segment is more than *Average(number of data in each segment)+ 2·Standard deviation(number of data in each segment)*, it determine that the segment is a burst.

Rattenbury et al. [3] gives the following method of using a Naïve scan to determine whether a tag is a time tag: For each segment i, let $T_r(x,i)$ be the usage count of tag x and $N_r(i) = \sum_x T_r(x,i)$. μN and σN respectively represent the average and standard deviation of $\left\{N_r(i)|i = 1 \cdots\right\}$. If $\frac{Tr(x,i)}{\mu N + 2\sigma N}$ is more than a certain threshold, the tag is determined to be a time tag.

3.2 Scale-structure Identification

The result of the Naïve scan method depends on the scale value r and segment separation determined by r. Hence, the selection of r determines whether a tag is a time tag or not.

To solve this scale problem, Rattenbury et al. presents a Scale-structure Identification method [3] that considers multiple scales based on Witkin's Scale-space method [5]. The Scale-structure Identification method defines scale values as follows: $R = \left\{r_k|k = 1 \cdot K, r_{k_1} > r_{k_2} \Longleftrightarrow k_1 > k_2\right\}$, R is selected such that that $r_k = \alpha^k, \alpha > 1.1$.

For each scale value r, do the following:

1. In T_x , consider the graph over T_x where the edges of a node pair exist if the distance between the two nodes is less than r.
2. Consider the set of the connected graphs (the set of clusters) Y_r, and calculate the entropy $E_r = \sum_{Y \in Y_r} \frac{|Y|}{|T_x|} \log_2 \frac{|T_x|}{|Y|}$.

E_r indicates the degree to which the whole distribution is similar to one cluster (When $|Y_r| = 1$, $E_r = 0$).

If the time distribution T_x of x has more than one dense region for all scales, the probability that the distribution is similar to one cluster is high. In other words, for all scales, the value of Er is likely small when the tag is a time tag. Thus, the method determines that the tag is a Time Tag when E_r is less than a certain threshold.

Figure 1 shows how the clustering changes for different scale values.

4 Extracting the concepts of the time and location tags

The concepts of the time or location tag can be extracted by using the Naïve scan or a variant of Scale-structure Identification that we describe below. For the sake of brevity, we describe only the case of an identified time tag.

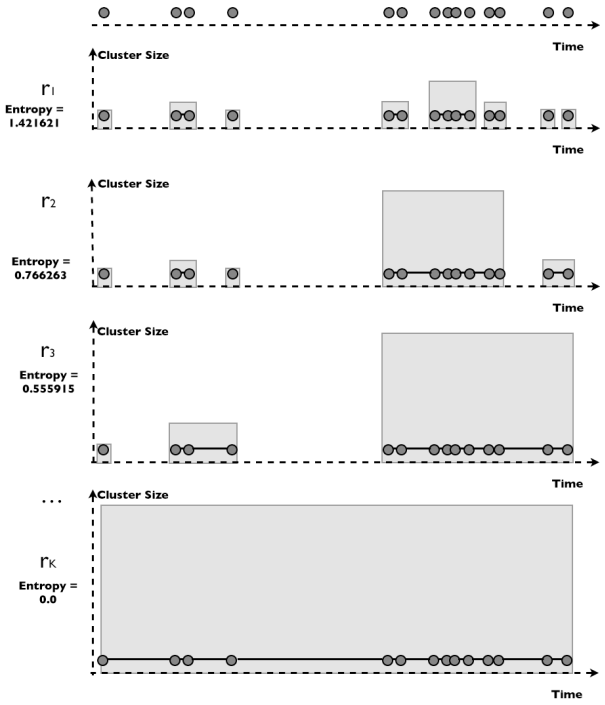


Fig. 1. An example of clustering

4.1 Naïve Scan

As described in Section 3, a Naïve scan determines that a segment is a burst if the number of data in the segment is more than a certain value. Hence, we can consider the segment is "the dense region" and the segment embodies the concept of a time tag.

4.2 Scale-structure Identification

The Naïve scan needs a pre-defined scale value and its results depend on this value: Thus, to get the best result, we divide a segment into multiple segments. To do so, we can employ a variant of Scale-structure Identification, which merges the results of analyzing clusters on all scale values. In order to extract the concepts of time tags, we have to solve two problems:

1. Which is the best clustering structure of scale value r to characterize the time distribution T_x ?
2. Once we know which clustering structure is the best, which cluster should be taken as embodying the concept of the time tag?

Selecting the clustering structure A better clustering structure has the following properties:

1. There is more than one cluster having a lot of nodes.
2. Clusters are distant from one another.

1 is needed because we want to specify the dense region of the distribution. If clusters are close, they could be combined. To represent the characteristics of the time distribution T_x, the Scale-structure Identification method calculates the entropy Er for each clustering structure, but entropy only indicates 1 not 2 because it does not consider the clusters' distance. Hence, we use compactness and isolation [6][7][8] as indications of 1 and 2. These measures are taken as evaluations of clusterings in the scale-space method:

$$compactness(Y_i) = \frac{\sum_{t \in Y_i} \exp^{-\|t-t_i\|^2/2r^2}}{\sum_{t \in Y_i} \sum_{Y_j \in Y} \exp^{-\|t-t_j\|^2/2r^2}}$$

$$isolation(Y_i) = \frac{\sum_{t \in Y_i} \exp^{-\|t-t_i\|^2/2r^2}}{\sum_{t} \exp^{-\|t-t_i\|^2/2r^2}}$$

t is each point in T_x , Y_i is the target cluster, Y is the set of tall clusters in the clustering structure, and t_i is the center point of Y_i . Compactness and isolation are related to a and b. Each value is less than 1.0, and bigger is better. Using the evaluation values of each cluster, we define evaluation formulas for the whole clustering structure [8]:

$$F_c(r) = \sum_{i}^{m} compactness(Y_i) - m$$

$$F_i(r) = \sum_{i}^{m} isolation(Y_i) - m$$

Here, m is the number of clusters in the clustering structure Y. The clustering structure is better if $F_c(r)$ and $F_i(r)$ are both bigger.

Characterizing the best clustering structure After selecting the best clustering structure, we determine which cluster best characterizes the clustering structure. If the number of nodes in a cluster is smaller than a percentage of all nodes, we consider that the cluster is characteristic; i.e., the cluster is the concept of the time tag. (Each tag can have multiple concepts.)

5 Evaluation

We implemented Naïve scan and Scale-structure Identification methods on actual Flickr data. We evaluated the following items:

time tags	august2007, carnival, comiket, cosmos, firework, fujirock, gameshow, ginkgo, gionmatsuri, jidai, june, newyear, obon, rama, september, snowboarding, tgs2006
location tags	beppu, chatan, daibutsu, enoshima, f1, hakodate, himeji, kanazawa, matsumoto, nara, otaru, roppongi, shinagawa, takayama, tokyodisneysea

Table 1. An example of time/location tags

1. Does the concept of the extracted time tag or location tag correspond to a concept identifiable by a human?
2. Does the concept extracted with Scale-structure Identification better correspond to human recognition than the concept extracted by the Naïve scan?
3. Is it possible to extract the tags' concept from the Flickr tag usage distributions?

5.1 Dataset

We collected photo data from Flickr, including photos taken between 2004/1/1 and 2007/12/31 and annotated with the location data of Japan. We excluded photo data whose upload date was before the taken date. On Flickr, the annotated location data to the photo was given an accuracy level between 1 and 16, with higher being more accurate. We removed photo data whose accuracy was less than 6.

We focused on tags annotating more than 100 photos and that were used by more than three users. The final data set contained 3,826,253 photos and 2,453 tags.

First, we manually determine that each tag is time tag or location tag and then randomly chose 100 tags from each set. Table 5.1 shows examples of time and location tags.

5.2 Experiment

We applied Naïve scan and Scale-structure Identification (SSI) to the selected tags. We analyzed multiple scales and selected the best scale by method and output the result by scale. We output the results of all scales to see if a suitable scale was chosen by the method. We use the scale values $r_k = 2, 4, \cdots$.

Each 100 time and location tags was manually ranked from 1 (low) to 5 (high). Figure 2 shows an example output, and Figure 3 shows an example rating.

The comparisons are:

- SSI: Scores for the results of the chosen scale by SSI.
- Best SSI: Scores for the best results of the SSI.
- Naïve Scan: Scores for the results of the Naïve scan.

The Naïve scan needed a pre-defined scale value so we calculated the average score for each scale and chose the best scale for it.

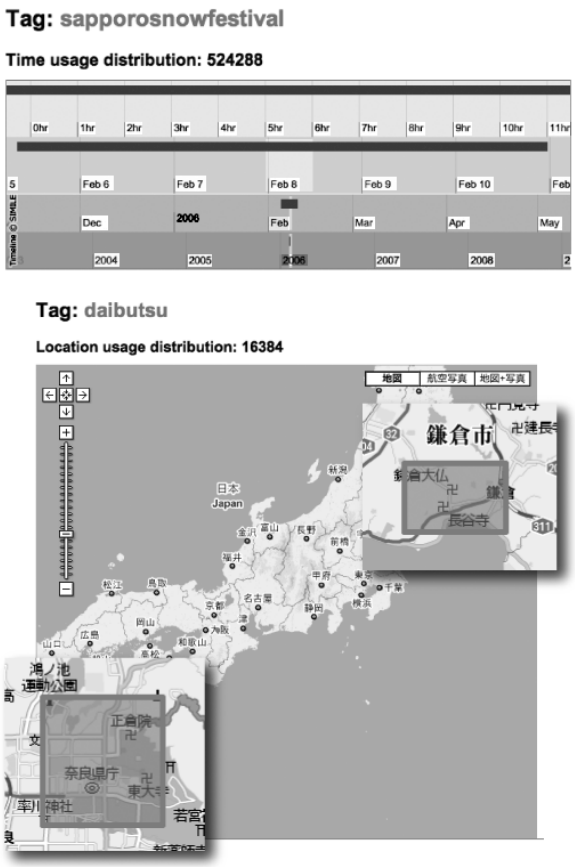


Fig. 2. An example of output

5.3 Result

Table 5.3 shows the average score for each method and the precision for right results when the score is higher than two. Figure 4 shows the distribution of differences between each method's scores and the best SSI scores. When the values are positive, the score for each method is better than the Best SSI. Lateral axis is the number of tags.

The average scores for SSI were near three, and precision was higher than 50%. These results confirm the concepts correspond to ones recognizable by humans. Average score and precision are higher for SSI than for Naïve scan. This confirms that our SSI method can get better results than the existing methods.

From Figure 4, it is often the case that the scores for Naïve scan are better (+1 to +4) or much worse (-2 to -4) than the best SSI scores; hence, the Naïve

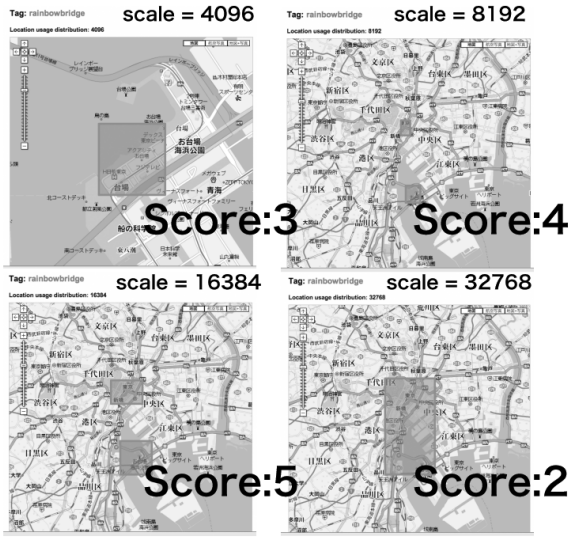


Fig. 3. An example of rating

		Average Score	Precision
Time	Best SSI	3.74	0.80
	SSI	2.70	0.56
	Naïve Scan	2.64	0.53
Location	Best SSI	3.86	0.84
	SSI	2.67	0.57
	Naïve Scan	2.62	0.49

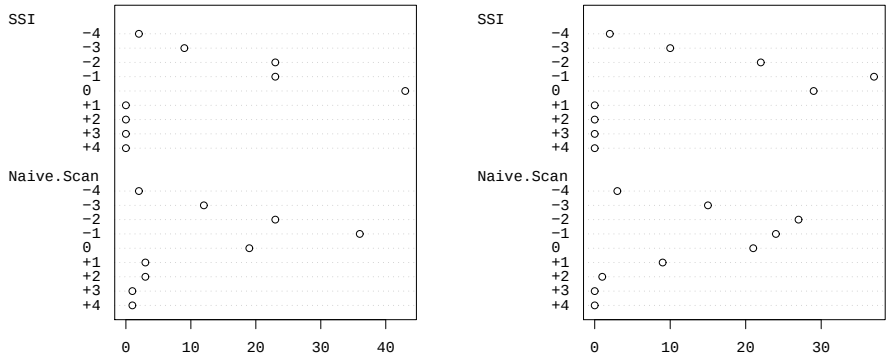
Table 2. Average score and precision for each method

scan's results vary in quality. Because SSI selects a suitable scale for each tag, its output is not as mismatched with human recognition as the Naïve method.

Therefore, the average best SSI score is near four and precision is higher than 80%. This means if we can select a suitable scale,we can extract the time or location concept of tags. Hence, our goal of extracting the concept of tags from their usage pattern is achievable.

6 Related Work

Researches about tagging system are well-practiced. Golder and Huberman [9] investigate how user do tagging and bookmarking on del.icio.us, a popular social bookmarking service. They discover regularities in user activity, tag frequencies and bursts of popularity in bookmarking. They also categorize the tags used for bookmarking into seven classes by their function. On the other hand, Marlow et



(a) Time tag (b) Location tag

Fig. 4. Distribution of difference between each method scores and Best SSI scores.

al. [10] identify taxonomy of tagging systems' design and user incentives. Moreover, they showed dynamics of Flickr and del.icio.us systems are quite different. Another research [11] survey the motivations for annotation and tagging photographs in mobile and online media, especially Flickr and ZoneTag [12]. [13] analyze tagging behaviour of users on Flickr and show the distribution of Flickr tags over the most common WordNet categories.

Improving usefulness of tags has been of increasing research interest. There is a work to assign specific schemes to facilitate interoperability between tagging systems [14]. Some methods are proposed to derive semantic structures from tags: obtaining semantic relations between the tags by using online ontologies [15], cluster tags by organizing undirected graph from tag space (Each node corresponds to a tag and Each edge is weighted related to co-occurrence frequency of a tag pair) [16] and deriving hierarchical semantics of tags by using unsupervised model [17]. Xu et al. [18] introduce a tag suggestion system. This system spot high-quality tags which determined by their popularity, coverage, etc.

Some research analyze the tags on Flickr using temporally information [19] or spatially information [20]. The system described in [19] generate the interesting tags on Flickr during specific time period by computing interestingness of each tags and visualize them. World Explorer [20] also focuses representative tags for each spatial region and obtain them by using techniques of multi-level clustering and TF-IDF based scoring.

There have been some previous works to extract semantic from Flickr tags, directly related to our work. Scale-structure Identification [3], as applied here, determines whether tags on Flickr are time and/or location Tags. We also adapted it to extract the concepts of such tags. Schmitz et al. [21] tried to extract an

ontology from Flickr tags using tag co-occurrence relations and organizing subsumption model and the research in [13] also analyzed the tag co-occurrence to build a tag recommendation system. Both works focuses on the tag semantics, similar to ours, but their approach is to extract synonyms from tag co-occurrence relations so the target tags' concept is different from ours.

7 Conclusion

In this paper, we focused on the new challenge of extracting the temporal and spatial concepts of tags from the tags' temporal/spatial usage distributions. To do so we modified the Scale-structure Identification, which is the existing method to determine whether a tag is related to a time or location, and showed that it can extract the temporal and spatial concepts of tags with higher accuracy than a Naïve scan in an experiment using actual Flickr data. We showed that if the method is able to select the suitable scale, it can extract the temporal or spatial concept of the tags with a high degree of accuracy, higher than 80%.

References

1. Vander Wal, T.: Folksonomy coinage and definition. http://www.vanderwal.net/folksonomy.html (2007)
2. : Flickr. http://www.flickr.com
3. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2007) 103–110
4. Vlachos, M., Meek, C., Vagena, Z., Gunopulos, D.: Identifying similarities, periodicities and bursts for online search queries. In: SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data, New York, NY, USA, ACM (2004) 131–142
5. Witkin, A.: Scale space filtering. Readings in Computer Vision: Issues, Problems, Principles, and Paradigms (1987)
6. Wong, Y.F.: Clustering data by melting. Neural Comput. **5**(1) (1993) 89–104
7. Chakravarthy, S.V., Ghosh, J.: Scale-based clustering using the radial basis function network. IEEE Transactions on Neural Networks **7**(5) (September 1996) 1250–1261
8. Leung, Y., Zhang, J.S., Xu, Z.B.: Clustering by scale-space filtering. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12) (2000) 1396–1410
9. Golder, S., Huberman, B.: The Structure of Collaborative Tagging Systems. Arxiv preprint cs.DL/0508082 (2005)
10. Marlow, C., Naaman, M., Boyd, D., Davis, M.: HT06, tagging paper, taxonomy, Flickr, academic article, to read. Proceedings of the seventeenth conference on Hypertext and hypermedia (2006) 31–40
11. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. Proceedings of the SIGCHI conference on Human factors in computing systems (2007) 971–980
12. : Zonetag. http://zonetag.research.yahoo.com/

13. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: WWW '08: Proceeding of the 17th international conference on World Wide Web, New York, NY, USA, ACM (2008) 327–336
14. Gruber, T.: Ontology of folksonomy: A mash-up of apples and oranges (2005)
15. Angeletou, S., Sabou, M., Specia, L., Motta, E.: Bridging the Gap Between Folksonomies and the Semantic Web: An Experience Report. Bridging the Gap between Semantic Web and Web Workshop at ESWC2006, Budva, Montenegro **2** (2006)
16. Begelman, G., Keller, P., Smadja, F.: Automated Tag Clustering: Improving search and exploration in the tag space. Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland (2006)
17. Zhou, M., Bao, S., Wu, X., Yu, Y.: An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotations. The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC (2007) 680–693
18. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the semantic web: Collaborative tag suggestions. Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland (2006)
19. Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., Tomkins, A.: Visualizing tags over time. In: WWW '06: Proceedings of the 15th international conference on World Wide Web, New York, NY, USA, ACM (2006) 193–202
20. Ahern, S., Naaman, M., Nair, R., Yang, J.: World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. Proceedings of the 2007 conference on Digital libraries (2007) 1–10
21. Schmitz, P.: Inducing ontology from flickr tags. Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, May (2006)

Case Sharing and Ontology Structuring in an Online Oral Medicine Community

Marie Gustafsson[1,2]

[1] School of Humanities and Informatics,
University of Skövde, SE-541 28 Skövde, Sweden
marie.gustafsson@his.se
[2] Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, SE-412 96 Göteborg, Sweden

Abstract. The Swedish Oral Medicine Web (SOMWeb) is an online system built to support knowledge sharing among oral medicine practitioners who hold monthly telephone conferences to discuss difficult and interesting cases. Semantic Web technologies are used to model the templates used for case entry, the ontology of values used in filling in cases, and community data. To study the practitioners' use and perceptions of the collaboration and the SOMWeb system, we have used observations of teleconferences, interviews with participants, and an online questionnaire. These are analyzed to provide an understanding of the participants' opinions about the structured case entry and why they do or do not contribute. This is followed by a discussion on future work where the value ontology is made available for community editing and structuring, and incentives for user contributions to this process.

1 Introduction

Oral medicine is a small but growing subdiscipline of dentistry, with geographically distributed practitioners. To enable distance consultations and promote learning, the Swedish Oral Medicine Network (SOMNet) has been holding monthly telephone conferences for over ten years, where difficult and interesting cases are discussed. In 2006, the Semantic Web-based Swedish Oral Medicine Web (SOMWeb) system [1] was introduced to support these meetings and case entry, browsing, and analysis. Use and perceptions of SOMWeb have been studied through interviews, observations, and a questionnaire. Before the introduction of SOMWeb, cases were e-mailed as PowerPoint-presentations among participants before meetings. With SOMWeb, user-defined templates are used to generate forms for entering data for different kinds of consultations. When filling in such a form, values that may be selected for each question are taken from a user-defined value ontology, to which the user may add a value if it is missing. In this paper, we present the users' thoughts on the structured case entry introduced with SOMWeb. We also outline future work on an online tool for structuring the value ontology and discuss its possibilities and limitations, especially with respect to incentives.

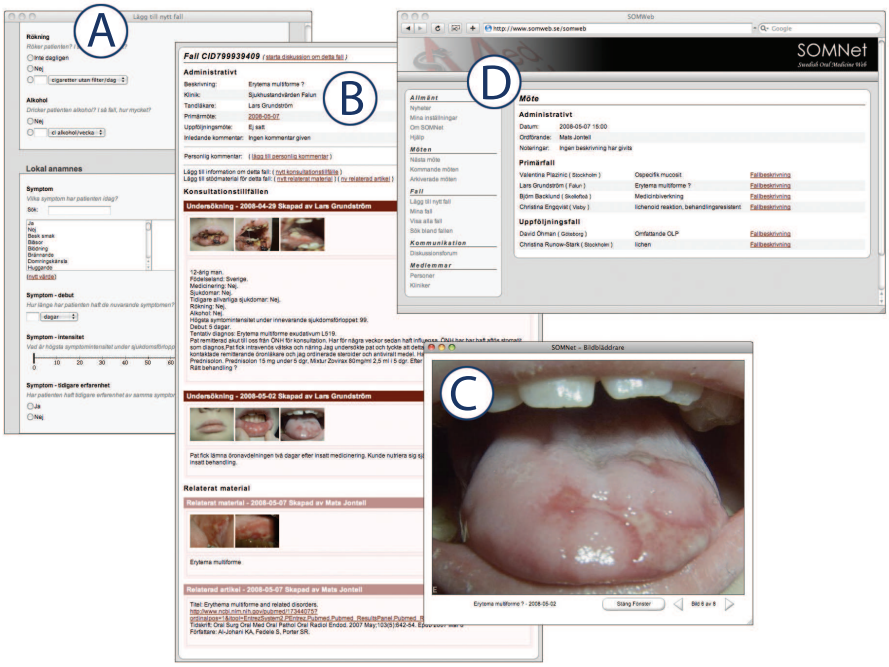


Fig. 1. The figure shows screenshots of key parts of SOMWeb: part of an examination data entry form (A), case presentation with pictures and text description generated from examination data (B), image browser (C), and a meeting page with case brought up for initial and follow-up consultation (D). All text is in Swedish.

2 The SOMWeb System

In September 2008, SOMWeb had 102 registered users located at 59 clinics. It has been used at 20 meetings and the case repository contains 105 cases. Currently, ten to fifteen clinics participate in each meeting. All members do not participate on each occasion, and there are meeting participants that are not SOMWeb members. The members are mostly dentists working in hospitals, primary care facilities, and private practice. Members provide their real names and workplace.

Before the development of SOMWeb began, meetings were observed and an online questionnaire was distributed. Several problems with the previous approach were identified, such as no shared record of discussed cases, that relevant information may be missing from the case presentation, and lack of written record of what was decided at the meetings. All these problems also make the use of entered case descriptions and results from meeting discussions as a basis for further analysis hard. The SOMWeb system was developed iteratively, including a selected group of users in the design process.

The functionality of SOMWeb is currently centered on cases and meetings. Figure 1 shows screenshots of important parts of SOMWeb. Meetings are added

to the system by users with an administrator role. Any member can enter a case and select a meeting for discussion. A link to the case presentation is automatically added to the page for that meeting. The owner can add more information about a case as it becomes available. All members can add information about relevant articles and information about related cases. The chairperson of a meeting can add case notes of what was suggested at the meeting. A user may also add private notes to any case. Cases in the repository can be viewed from the meeting pages, a list of all cases in the system, and via free text search.

The structured case entry form is generated from a user-defined OWL template of the examination. A template consists of categories (e.g., `PatientData` and `MucosAnamnesis`), with associated questions (e.g., current symptoms of the patient). The values that may be used in answering questions are instances of classes (e.g., `Diagnosis` and `Allergy`) in a value list ontology. The value list ontology was generated from a previous system of the research group, and all lists were initially flat, i.e., there were no subclasses of e.g., `Diagnosis`. Individual cases are stored in RDF. When viewing the case presentation, a case summary is generated from the RDFS labels. OWL is also used to model community aspects such as users, meetings, and cases, and data related to these are stored in RDF.

3 Methods for Studying Participants' Use and Perceptions

As part of a larger effort to study the use of SOMWeb and the communication of SOMNet, we have used an online questionnaire, interviews, and meeting observations. The online questionnaire had both open-ended and closed-ended questions, including a comparison of the SOMWeb system with the previous PPT-based approach. Responses were collected during one month in the spring of 2007, and 24 out of the at the time 60 members responded. From late 2007 to early 2008 nine members of SOMNet were interviewed to increase our understanding of how SOMWeb is used and of how it has affected SOMNet. The semi-structured interviews included questions on how the members perceived the new method for entering cases and the values list used. Of the interviewees, three had been members more or less from the start, three had been members for at least four years, and three had joined more recently. Each interview lasted between 35 and 85 minutes. Ten teleconferences have been observed by sitting at five different clinics during the meeting. These were carried out with the aim of seeing how cases were presented, how the participants behave locally, and how the system is used locally during meetings.

4 Case Entry

In reply to the questionnaire, 88 % found viewing old cases better in SOMWeb, and 12 % were neutral. Of the 24 persons answering the questionnaire, 29 % had added cases. Of these, 87 % thought adding cases was better in SOMWeb,

and 13 % were neutral. Interviewees stated that SOMNet's collaboration has improved with the SOMWeb system. Motivations were e.g., easier and less time-consuming case entry, more uniform case data, and the collected view of a case over time. Of the interviewees, six out of nine have added cases. Four find case entry easier with the new system compared to using PowerPoint. Two had difficulties: One used only the free text entry of the form, finding that it took too much time to fill in the form. The other brought up difficulties in deciding which data to enter for patients with complex clinical situations. An interesting conflict was identified where one interviewee thought duplicate and misspelled entries in the value list were problematic, while others found the breadth of values good and believed it impossible to have lists with no odd values. One interviewee thought questions were missing from the form. There is a tool in the system where administrators can upload new examination templates, which has not yet been used. It is probably the case that the community does not yet have processes in place to handle this issue and the current template is "good enough". While one respondent had areas of interest that they wanted to be included, others voiced concern that they form would become too long.

From the observations and interviews we see mainly three purposes for adding and presenting a case: seeking advice regarding diagnosis or treatment, unusual cases, and where the presenter wants to raise an issue for discussion. Seeking advice is most common. About 25 % of the members have submitted at least one case. Of the 105 cases in the repository, five people have submitted about 50 %. One person has submitted 20 cases, which may be attributed to chairing (the chairpersonship of the meeting rotates among a several active members) meetings where few cases had been entered. There have been discussions among the core members of the group of how to get those less active to add cases and to speak at meetings. They have speculated that one issue is concern over revealing gaps in one's knowledge. Some replies to the questionnaire, upon the question if they had considered adding a case but had not, indicated worry that it was not "advanced enough". It has been suggested that one way to alleviate this is for senior members to add straightforward cases. Further, contrary to the worries of junior members, the experts find that what appear to be straightforward cases often lead to interesting discussions. Finally, a lack of time was an issue often raised by participants, either due to a heavy load of patients or teaching. This indicates the importance of easy to use tools.

5 Community Ontology Editing and Structuring, and Incentives for User Contributions

The interviewees find that as the number of cases in the system increases, more advanced methods of browsing and searching the cases are needed. One way of providing this is by adding more detail to the ontology from which instances are selected in entering case data. The current value list ontology contains no subclasses of e.g., `Diagnosis`. We are therefore interested in providing a tool to let the users provide more structure and detail to this ontology to enable

improved exploration of case data. In developing such a tool, there are several concerns. A major one is how to motivate users to contribute to the ontology structuring. Another is how to accommodate different conceptualizations of the domain.

In our research group's previous work to support oral medicine practitioners, a data analysis tool was developed. In the tool, the user may create aggregates of values to be used in grouping data, e.g., diagnosis categories. These aggregates are taken as a starting point for a more fine grained ontology for use in the case browser, but they do not cover the whole new value list, and some aggregates have been created with a certain analysis task in mind. Since diagnosis subcategories are well-covered by the aggregates, the users can use the case browser to get a more detailed view of subgroups of diagnoses. Through using this tool, a user may then discover that certain values are missing from e.g., a diagnosis subclass, and needs to be given the opportunity to add the value.

We are also considering adding a separate tool to SOMWeb to make groupings of values (subclassing) to be used in the browser. Initially the user may only want to "scratch their own itch", but that they can then decide to make the grouping public. This approach would both increase benefit for the structure provider, as well as permitting the user to create and test it in a way that does not lead to apprehension of exposing gaps in one's knowledge. A drawback of this approach is of course that users may opt mostly for the private approach.

6 Discussion

The purpose of introducing structured case entry is to attempt to gather all relevant data for cases. An immediate benefit of this is that this data is at hand for meetings. Further, it makes possible the case browser tool described above, which is currently under development. However, we also view structured case entry as a prerequisite for learning from clinical data. Our study of the use and perceptions of SOMWeb have lead us to find that its users enjoy the collaboration and find it useful, have slightly different opinions of on the goals of this collaboration and how it should be carried out, but agree that more people should be encouraged to participate and that lack of time is a barrier to most members. That only 25 % of members have submitted cases can be compared with the findings of Nonnecke and Preece [2] that lurkers often make up at least half of the subscribers of discussion lists. If we look at the reasons that two interviewees found the structured case entry unsatisfying (see Sec. 4), we see an inclination for narrative and reservations with distilling a patient's case to the structure of the form. While it may be possible to alleviate such issues with e.g., another interface, it also points to more general problems in deciding between structured data versus a narrative form. Related to this is the trade-off between completeness and complexity. If a more detailed form was provided, or maybe different forms for different diagnoses, then a more complicated clinical situation could be captured. However, filling in such a form would be more time-

consuming, which is also the case if more questions are added to cater to different interests.

Siorpaes and Hepp [3] observe that in ontology building the effort and benefits are often separate. In an approach where the structuring is done to perform analyses relevant to the user, some of this may be overcome. Another issue that often arises with knowledge sharing is that of trust, and such is the case here as well. For example, there has to be trust in the structures provided by others, and participants must trust that their contributions are taken seriously. Connected with trust is provenance, in this case knowing who contributed e.g., a new class to the ontology. This makes it possible to trace thoughts and find explanations for added structures. The creation of trust is a complex psychological and sociological issue. We believe that persons in a community with leadership roles are important in creating and maintaining trust in the community process and products. Thus, these people will probably be central in the structuring of the SOMWeb ontology. This may also be gleaned from that five members have contributed 50 % of the cases. One may also observe that certain people more quickly take on a curator role, and maybe such a role should be provided in addition to the administrator role. In our interviews, for example, it became apparent that the respondents have rather varying sensitivities to detail. These differences must be handled in the tool as well, though perhaps they should be seen as a possibility rather than an issue, in that certain people will be more apt to perform clean up activities. White and Lutters [4] discuss the difficulties in getting heterogeneous groups to agree on a view of a subject and the level of granularity that should be used. This may be the case in SOMNet as well, and it will then have to be decided whether several conceptualizations shall be seen as valid or whether there should be a group process to decide upon one conceptualization.

Acknowledgements

Thanks to Göran Falkman, University of Skövde, Olof Torgersson, University of Gothenburg, and Mats Jontell, Sahlgrenska Academy, University of Gothenburg, for comments. This work is funded by the Swedish Governmental Agency for Innovation Systems (VINNOVA), research grant 2006-02792.

References

1. Falkman, G., Gustafsson, M., Jontell, M., Torgersson, O.: SOMWeb: A Semantic Web-based system for supporting collaboration of distributed medical communities of practice. J Med Internet Res **10**(3) (2008) e25
2. Nonnecke, B., Preece, J.: Lurker demographics: Counting the silent. CHI Letters **2**(1) (2000) 73–80
3. Siorpaes, K., Hepp, M.: Games with a purpose for the Semantic Web. IEEE Intelligent Systems **23**(3) (2008) 50–60
4. White, K.F., Lutters, W.G.: Structuring cross-organizational knowledge sharing. In: Proc. GROUP '07. (2007) 187–196

Bridging the Motivation Gap for Individual Annotators: What Can We Learn From Photo Annotation Systems?

Tabin Hasan[1] and Anthony Jameson[2] ⋆

[1] University of Trento
Trento, Italy
[2] Fondazione Bruno Kessler
Trento, Italy

Abstract. The importance of incentives and socially based motivation for metadata generation should not distract attention entirely from the need to design tools for metadata generation that use every means available to maximize the efficiency and intrinsic motivation of the individual annotator. The popular application domain of (individual) photo management has recently given rise to a number of strategies and methods that can serve as a source of inspiration for the design of metadata generation support for the semantic web. This position paper offers a brief synthesis of relevant work that is intended to serve as a basis for the representation of this perspective at the Insemtive 2008 workshop.

1 Why Photo Annotation Is a Relevant and Instructive Scenario

The problem of motivating contributions to a community-supported resource (of which the semantic web can be seen as an especially ambitious example) is often framed in terms of a contrast between the interests of an individual contributor and the interests of the group as a whole (see, e.g., [1]): If only people were as motivated to contribute to the semantic web as they are to their own personal knowledge bases, it would seem, the creation of metadata for the semantic web would thrive.

While this perspective is valid and important, we would like to call attention to the fact that there can also be a major "motivation gap" when individuals are making similar contributions for their own benefit. Consequently, we also need to examine ways of closing the motivation gap that arise even when individuals are working for their own benefit. These methods can in turn also benefit the community-supported semantic web indirectly.

More concretely, consider the familiar problem of adding metadata to photos: Since photos form a natural part of the semantic web as well as of many Web 2.0 systems, improving people's motivation to add metadata to photos would constitute a contribution to the goals of this workshop. But even when an individual is managing their own personal photo collection, there is a challenging motivation gap: Having good metadata would make it much easier for the user to accomplish common tasks such as searching

⋆ The research described in this paper is being conducted in the context of the KnowDive project (http://disi.unitn.it/∼knowdive). The participation of the second author was supported by Vulcan, Inc.

for photos that fit a particular description; but as has often been noted (see, e.g., [2]), few users get very far in adding such metadata, largely because of the time-consuming and tedious nature of the work that is involved.

Because of the rapidly growing popularity and practical importance of digital personal photo collections, a good deal of research has been devoted in recent years to the problem of motivating and/or supporting untrained end users in adding metadata to their photos. Despite—or indeed because of—the differences between this scenario and the more general scenario of adding metadata for the semantic web, it is worthwhile to look closely at the successes that have been achieved in this area and to consider how they might be generalized.

2 Overview of Determinants of Successful and Motivating Photo Annotation

Types of metadata that users often want to add to photos include (a) persons, objects, locations, and events depicted in the photos; and (b) information about the context in which the photo was taken (e.g., "just before sundown" or "just after the end of the championship football game"). It is often assumed that the photos already have accurate time and location stamps that can serve as input to automatic processing (though in fact such automatically generated metadata may be missing or incorrect for various reasons and may therefore need to be supplied by the user—a problem to which some of the metadata creation approaches discussed below can be applied).

Figure 1 summarizes a number of the ideas that have emerged from recent work on interfaces that help users to add such metadata. Before discussing these points individually and illustrating them with reference to recent research, we will comment on them briefly.

In terms of motivation, the overall approach taken in photo annotation systems for individuals is not based on external incentives or social mechanisms but rather on the provision of an intrinsically motivating experience for the individual user. Somewhat more concretely, the strategy is to optimize the relationship between (a) the cost to the user in terms of work done (in particular, tedious work) and; (b) the benefits in terms of enjoyable experiences, successful task performance, and visible improvements to the collection of items.

In some ways, the most straightforward approach is to exploit *external resources* (see the bottom left-hand corner of the figure) that can straightforwardly generate new metadata on the basis of existing metadata (e.g., supplying the name of a town on the basis of GPS coordinates). But external resources may also serve as input to sophisticated *algorithms* that analyze the content of items, either suggesting metadata or at least grouping together items that appear (to the system) to belong in the same category. Since such algorithms do not in general perform perfectly, there is generally a *user interface* that is designed to enable the user to supply the necessary manual input with minimal effort and maximal enjoyment. The *user input* itself can be seen as a valuable resource, which includes both explicit *annotation* actions and *naturally occurring* actions that provide useful information although the user does not perform them specifically for the purpose of adding metadata.

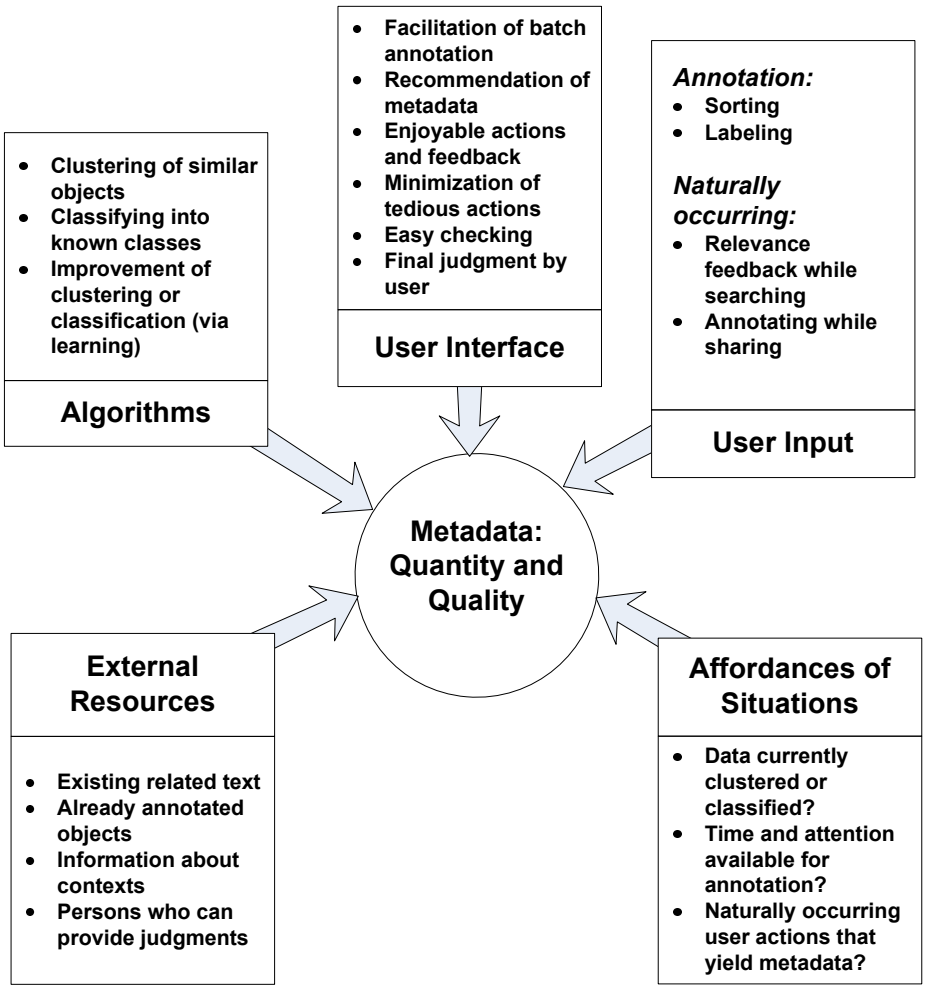


Fig. 1. Overview of factors that can contribute to the quality and quantity of metadata added in a sophisticated system for the individual annotation of resources such as photos.

Finally, some systems take into account and exploit the *affordances of situations*, taking into account the fact that people use their photo management systems in a variety of situations, each of which offers certain possibilities and limitations in terms of metadata generation.

As we will see in the next sections, these five contributors to metadata generation do not contribute independently in an additive manner. Often, a favorable combination of two or three contributors is required to achieve good results. For example, a classification algorithm may work well only on the basis of information in an external database; and it's output may be manageable only with a cleverly designed user interface that

elicits the necessary user input with minimal effort in an especially favorable situation. One objective of this position paper is to encourage this holistic view of the various contributing factors, whereas most of the primary research literature understandably focuses on one or two factors.

We will now briefly discuss some representative examples of systems that illustrate the contributing factors shown in Figure 1.

3 External Resources

Naaman et al. ([3]) provided a relatively early demonstration of how a variety of types of contextual metadata can be added to geo-referenced digital photos with the use of off-the-shelf and web-based data sources. The types of metadata added included the local daylight status and the local weather conditions. In addition to showing the feasibility of automatically adding contextual metadata, the authors showed how such metadata can be useful for searching and browsing, despite the fact that they may seem at first glance not to be especially important. For example, when searching for a given photo people may have a hard time characterizing the content of the photo itself yet find it easy to characterize the weather and daylight status—which may together narrow down the search space dramatically. A lesson for semantic web metadata creation is that the intrinsic importance of the metadata should not be the only criterion for deciding whether they are worth adding.

Another well-known system that uses this approach is PHOTOCOPAIN ([4]). This system also illustrates how an external resource can be used to support a sophisticated algorithm: Tagged photos on flickr.com serve as training data for the system's image analysis algorithms.

4 Algorithms and User Interfaces

A compelling example system in which algorithms play a central role is SAPHARI ([5]). One of the algorithms uses the clothes worn by people in photos for the heuristic clustering of photos that presumably depict the same person. This approach is an example of the clever exploitation of the strengths of the computer and the human, respectively: The computer does the tedious work of putting into a single place all of the photos that show a person wearing a particular set of clothes; all that remains for the user is to check whether these photos do in fact depict the same person and to supply the identity of that person. Note that the output of the algorithm would be useless if it were not combined with a suitable user interface.

Automatic photo clustering is also done in the EASYALBUM system ([6]), here on the basis of the similarity of faces or scenes. The results of the clustering are exploited in subtle ways throughout the interface—for example, in order to minimize the amount of scrolling that is required.

Some systems that provide clustering or classification algorithms also provide machine learning mechanisms that boost the performance of the algorithms over time for a particular user or collection. For example, whenever EASYALBUM (mentioned above)

receives new user input indicating the correct annotation of a given photo, the performance of the clustering algorithm is adapted accordingly. An approach that is apparently still new with regard to photo annotation systems for end-users is *active learning* ([7]; [8]): The system attempts to minimize the amount of input required of the user by determining at each point in time which additional training examples would be most helpful.

5 User Input

We have already seen several strategies for minimizing the number of explicit annotation actions required of the user by allowing the system to make maximal use of each such action. A different approach to optimizing the use of the user input is to interpret actions that involve no (or minimal) additional effort on the part of the user beyond the effort that they would normally exert in performing non-annotation tasks with their photo management system.

For example, in MIALBUM ([9]), a search algorithm for photos is made available that includes the opportunity for the user to supply relevance feedback by explicitly indicating which of the photos returned for a given query in fact satisfy the query. This relevance feedback is then used as input for enhancing the metadata associated with the photos in question. Given that relevance feedback is in principle worthwhile even just in terms of improving the results of the current search, its exploitation for metadata enhancement can be seen as not requiring additional user effort.[3]

Other types of natural user action that can be exploited include actions that occur when the user communicates with other persons about the photos in the collection—for example, when sending photos to another person ([10]) or or when discussing photos with other persons face-to-face (see, e.g., [11]).

6 The Affordances of Situations

The examples just mentioned illustrate the more general points that (a) photo annotation systems are used in a variety of settings and (b) each such setting typically offers some particularly good opportunities for metadata generation (as well as being limited with respect to other types of metadata generation). It therefore makes sense to design an annotation system so that it can exploit the specific potential (or *affordances*, to use the term from the HCI literature) of each situation. To take a simple example: When a user is uploading photos from their camera's memory chip, there is a good chance that many or all of the photos concern a single event (e.g., a wedding or a vacation). Moreover, at this point in time the user is likely to have a relatively precise recollection of the relevant facts. This is therefore an especially favorable time to encourage the user to make bulk annotations: Once these photos have flowed into the ocean of already stored photos and the relevant events have faded in the user's mind, adding the same metadata would present more of a challenge for both the system and the user.

[3] The authors point out that, in reality, getting users to supply relevant feedback is still a partly unsolved interface design problem, despite the immediate utility of such feedback.

7 Concluding Remarks

If you want to motivate a person to mow their lawn every week, you can offer some material incentive or set up a social mechanism by which they earn approval if they mow their lawn and perhaps disapproval if they fail to do so. A different approach is to take away their clumsy mechanical lawn mower and give them a well-designed and -engineered electric mower that makes it fun and intrinsically rewarding to mow the lawn in just a few minutes.

Strategies of the first type will presumably attract the most attention in the Insemtive 2008 workshop, and they certainly are important for the semantic web. Our position is that such approaches work best when combined with approaches of the second type; and that many generalizable ideas along these lines have recently emerged that have not yet made it into the mainstream literature on metadata generation for the semantic web.

References

1. McDowell, L., Etzioni, O., Gribble, S.D., Halevy, A., Levy, H., Pentney, W., Verma, D., Vlasseva, S.: Mangrove: Enticing ordinary people onto the semantic web via instant gratification. In: Proceedings of ISWC 2003, Sanibel Island, Florida (2003) 754–770
2. Rodden, K., Wood, K.R.: How do people manage their digital photographs? In Terveen, L., Wixon, D., Comstock, E., Sasse, A., eds.: Human Factors in Computing Systems: CHI 2003 Conference Proceedings. ACM, New York (2003) 409–416
3. Naaman, M., Harada, S., Wang, Q., Garcia-Molina, H., Paepcke, A.: Context data in geo-referenced digital photo collections. In: Proceedings of the Twelfth International Conference on Multimedia, New York (2004) 196–203
4. Tuffield, M.M., Harris, S., Dupplaw, D., Chakravarthy, A., Brewster, C., Gibbins, N., O'Hara, K., Ciravegna, F., Sleeman, D., Shadbolt, N., Wilks, Y.: Image annotation with Photocopain. In: Proceedings of the First International Workshop on Semantic Web Annotations for Multimedia, held at the World Wide Web Conference. (2006)
5. Suh, B., Bederson, B.B.: Semi-automatic photo annotation strategies using event based clustering and clothing based person recognition. Interacting with Computers **19** (2007) 524–544
6. Cui, J., Wen, F., Xiao, R., Tian, Y., Tang, X.: EasyAlbum: An interactive photo annotation system based on face clustering and re-ranking. In Begole, B., Payne, S., Churchill, E., Amant, R.S., Gilmore, D., Rosson, M.B., eds.: Human Factors in Computing Systems: CHI 2007 Conference Proceedings. ACM, New York (2007) 367–376
7. Zhang, C., Chen, T.: An active learning framework for content-based information retrieval. IEEE Transactions on Multimedia **4**(2) (2002) 260–268
8. Cord, M., Gosselin, P.H., Philipp-Foliguet, S.: Stochastic exploration and active learning for image retrieval. Image and Vision Computing **25** (2007) 14–23
9. Wenyin, L., Dumais, S., Sun, Y., Zhang, H., Czerwinski, M., Field, B.: Semi-automatic image annotation. In: Proceedings of Interact 2001, Eighth IFIP TC.13 Conference on Human Computer Interaction. (2001)
10. Lieberman, H., Rosenzweig, E., Singh, P.: Aria: An agent for annotating and retrieving images. IEEE Computer (2001) 57–61
11. Barthelmess, P., Kaiser, E., McGee, D.: Toward content-aware multimodal tagging of personal photo collections. In: Proceedings of the Ninth International Conference on Multimodal Interfaces. (2007) 122–125

Growing the Semantic Web with Inverse Semantic Search

Hans-Jörg Happel

FZI Research Center for Information Technology, Karlsruhe, Germany
happel@fzi.de

Abstract. Many use cases for the Semantic Web assume the availability of public metadata. However, research has not yet addressed in a satisfactory manner why and how metadata is published on the Semantic Web. We analyze several reasons and barriers for creating and sharing semantic metadata. In particular, we address the issue of how metadata from private spaces can diffuse into the public Semantic Web. Therefore we introduce the concept of inverse semantic search – an approach which aggregates information needs to motivate information providers to share private metadata.[1]

1 Introduction

The vision of the Semantic Web [1] describes a web populated by machine-understandable metadata based on which agents can reason and act to fulfill tasks for human users. However, the realization of the Semantic Web largely depends on the availability of such structured metadata.

While the usefulness of metadata has been claimed for many domains and applications, publicly available metadata in the Semantic Web is still scarce. Two main issues impede a widespread success of metadata [2]. First, metadata is additional, descriptive data on top of actual information resources by its very nature. Thus, it is not created for self-purpose but costs additional effort. Secondly, the creation of metadata often implies a disparity of providers and beneficiaries (i.e. people using metadata are different from people creating it) and between the time of creation and its use [3].

While a number of studies have investigated the forces that drive the creation of metadata by individual users (mostly w.r.t. tagging systems, c.f. section 2.2), there exists no unified theory why semantic metadata is created and how it is made available [2]. Especially the Semantic Web vision does not address the *creator* side of metadata, but focuses on the consumer side and its applications. This is quite similar to the domain of information retrieval, which also neglects the role of information *providers*.

Within this paper, we will line out an initial theory about why and how metadata is created and thus how the Semantic Web could be populated. We

[1] This work has been supported in part by the TEAM project, which is funded by the EU-IST programme under grant FP6-35111 and the BMBF-funded project WAVES.

therefore analyze different aspects of metadata in the following section. Based on those insights, we present the conceptualization and realization of our approach called *inverse semantic search*, which guides potential metadata providers using aggregated information needs. We claim that the design of inverse semantic search thus provides motivational incentives to help growing the Semantic Web.

2 On metadata

2.1 Usage of metadata

We distinguish two major scenarios that motivate the usefulness of metadata in the Semantic Web. The most prominent one is *resource description for information retrieval*. The need for metadata in this scenario stems either from resources which are not accessible by standard keyword-based search (i.e. photos or videos), or from the fact that resources might not contain certain keywords/conceptualizations by which they might be accessed. Metadata is thus added to provide descriptive information which can incorporate structured classifications (like in library catalogues) or synonym keywords.

The second case for metadata is rooted in *task automation*. This comprises a whole range from visionary agent-driven scenarios which automatically perform actions on behalf of their human owners down to mash-ups where data from different sources is joined to provide some extra functionality [4].

2.2 Creation of metadata

We distinguish three different ways of creating metadata: 1) either it comes for free and just needs to be exposed, 2) it can be generated automatically or 3) has to be created manually.

The *exposition* case is the most simple one. If data is already available in some highly structured form, such as in database systems, it can easily be exposed. An example for this could be a cinema which offers metadata about available films out of its existing booking system. Although supporting tools already exist (e.g. [5]) an initial technical investment might be necessary to make such data available for external users.

The *automatic creation* of metadata tries to generate descriptive metadata using certain algorithms. Typical examples are machine learning systems which analyze documents, pictures or other content to automatically assign topics or categories. Such techniques depend on the availability of sophisticated algorithms, suitable input and training data and suffer from potential impreciseness [6]. Furthermore, they can not create arbitrary metadata (e.g. movie ratings or reviews). Automatic metadata creation techniques are therefore often used semi-automatically to assist human metadata creators.

Despite of its cost, *human created metadata* is thus still an important issue. While human metadata creation has been common for specific tasks such as library management, it has seen a renaissance in recent years due to the emerging

Web 2.0 phenomenon. Applications like del.icio.us[2] or Flickr[3] collect small pieces of metadata from individual users and unfold their power by aggregating them.

Motivational issues have been discussed concerning tagging and photo sharing systems in recent years. Results highlight the important role of personal and social benefits as functional motivations [3, 7, 6]. However, tagging systems can not be directly compared to general metadata for the Semantic Web. The authors of [2] discuss motivations for metadata sharing on a more abstract level, identifying advertising and retrieval services as potential contributors.

2.3 Visibility of metadata

Even if metadata has been created and is in place, it needs to be available for all its potential consumers. Like any kind of digital resource, metadata can be kept in arbitrary spheres of access – ranging from the private sphere of an individual user up to public visibility in the internet.

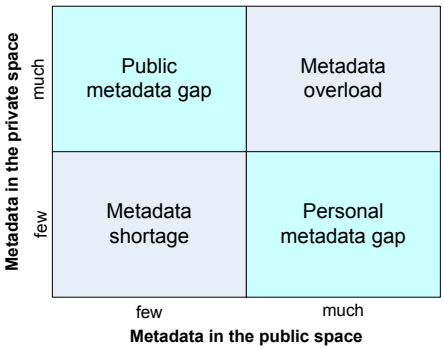


Fig. 1. Possible distributions of metadata in private vs. public information spaces (adapted from [8])

Private spheres are commonly used because users often hesitate to share data openly. Reasons are low motivation due to a lack of personal benefit [9–11], privacy concerns [12, 13] and effort for sharing (e.g. capturing, categorization and setting access rights) [13, 14]. Thus, even many open "Web 2.0" applications such as Flickr or del.icio.us allow for storing metadata privately.

Figure 1 illustrates this situation. It distinguishes the amount of metadata available for a certain information resource in the private space of a particular user vs. the public space. Four general situations are depicted: in a balanced situation, there either exists few metadata (*Metadata shortage*) or lots of metadata (*Metadata overload*) in both, the private and public space. If there is more

[2] http://del.icio.us
[3] http://www.flickr.com

metadata in the public space than in the private space, we call this a *personal metadata gap*. The case of a *public metadata gap* describes that no or only few metadata concerning an information resource exists in the public space, but in the private space of at least one particular user.

When considering the Semantic Web, the situations of a *public metadata gap* and *metadata shortage* are the most unfortunate ones, since potentially useful metadata is hidden in private spaces or does not exist at all.

2.4 Conclusion

Contrasting this section with the vision of the Semantic Web, the fact that the Semantic Web so far neglects the perspective of metadata *providers* has two consequences:

- The creation of metadata should be guided resp. focused, since it is a costly process.
- Feedback channels and easy sharing facilities should be incorporated in Semantic Web tool design.

In the following section, we will describe a general framework and an approach called *inverse semantic search* which tries to address incentives for sharing and creating semantic metadata.

3 Inverse semantic search

In this section we will describe a concept called *inverse semantic search* in order to help growing the amount of metadata in the Semantic Web. Therefore, we differentiate between *consumers* and *providers* of semantic metadata.

We will begin the section with a motivating scenario for our approach, followed by a specificiation of requirements and use cases. Afterwards, we will describe the realization of *inverse semantic search* in terms of its architecture and process steps.

3.1 Motivational example

As-is situation Our scenario involves two persons: Chrissy, who wants to buy a birthday present for her boyfriend, and Dave, who is a movie enthusiast. Chrissy's initial idea is to buy a trip to one of the locations mentioned in the movie "Casablanca" of which her boyfriend is a big fan. Thus, Chrissy queries her favourite Semantic Web search engine for "All locations mentioned in Casablanca". To her surprise, the application only returns the obvious "Casablanca" as a result – no additional metadata seems to be available on the web. However, Dave maintains his own local movie application, where he keeps data about his favourite films. His application actually contains "Paris" and "Lissabon" as additional locations mentioned in Casablanca. Since this data is within Dave's private space, Chrissy is not able to retrieve that information. Thus, she finally decides to buy a different birthday present.

To-be situation In order to improve knowledge sharing in the described situation, we propose that Chrissy's query is not just matched against the available metadata corpus (yielding only one result in our example), but also stored in a central query log. This information can then be made available to interested clients. Thus, Dave's movie application can retrieve this list of queries and automatically compare it to the metadata in his private space. In our example, this would reveal that information from Dave's computer could help satisfying Chrissy's information need. The movie application would present a list of metadata items to Dave, indicating that there is an information need that can be satisfied by sharing them. Dave may then choose to contribute this metadata to the public space. Once Dave shares the information, Chrissy could be notified about the new results.

Clearly this is a rather simplified example, which could probably be solved without any Semantic Web technologies at all. However, it illustrates the key principle of sharing and matching information needs asynchronously which is also applicable to scenarios utilizing more structured metadata.

3.2 Specification

In this section we specify our envisioned functionality by introducing a number of use cases and non-functional requirements.

Requirements As lined out in section 2, metadata provision suffers from a number of barriers. We want to address these barriers by satisfying the following set of non-functional requirements:

R1. Retain privacy An information provider must not expose information to others by default. Knowledge sharing systems often lack acceptance, since contributing information to the public space means losing control about it. However, many information providers want to retain such control, since information might be premature or sensitive [15].

R2. Minimize effort The effort for both, information providers and information seekers should be minimized. There should not be much redundant information provision [6].

R3. Motivate to share Information providers should be motivated to share relevant information with information seekers. Traditional knowledge sharing applications usually require to share information without signaling any benefit to the provider. Thus, those practices are often perceived as self-purpose with an unclear value. In opposite to this, we want to give the potential information provider more concrete information that can help to estimate the benefit of sharing certain metadata. Research targeting movie rating systems has shown that design features motivated by social psychology such as highlighting the uniqueness [16] or value [17] of a contribution can significantly increase information provision.

Use cases Metadata is typically queried in structured query languages such as SPARQL [18]. For the scope of this paper, we restrict ourselves to a fragment which allows to query for either *instances* or *literal values*.

Information need	Informal	Semi-formal
L	Chrissies Phone number	?x: ns:Chrissie ns:phoneNumber ?x
L	Speed of all cars	?x: ?y rdf:type ns:Car . ?y ns:hasSpeed ?x
I	Locations mentioned in Casablanca	?x: ?x ns:mentionedIn ns:Casablanca
I	All movies	?x: ?x rdf:type ns:Movie
I	Chrissies birth town	?x: ns:Chrissie ns:bornIn ?x
I	All videos tagged with "Chrissie"	?x: ?x ns:hasTag "Chrissie"
I	Places where Popes were born	?x: ?y ns:bornIn ?x . ?y rdf:type ns:Pope
I	All persons that own a car	?x: ?x ns:owns ?y . ?y rdf:type ns:Car

Table 1. Use cases (*ns* stands for an arbitrary namespace)

Table 1 shows example queries to illustrate eight different kinds of triple patterns which we consider in this paper. The first column shows the type of information need (instances or literal). The second column contains a written description of the information need. In the last column, a simplified formal representation of these information needs is shown. It contains the queried variable (*?x*) followed by constraints on this variable. Constraints are either concrete values for object or datatype properties (e.g. *?x hasTag "Chrissie"*) or types of object property values (*?x rdf:type ns:Movie*).

3.3 Realization

In this section we give a short definition of our approach. We then discuss architectural implications and describe the process steps involved.

Definition As already lined out, common retrieval models follow a *provide first – retrieve then* approach. This means that they do not conceptualize the provision of information but assume that information exists at the time of retrieval. Information seekers can then query this information to retrieve results satisfying their information need.

The basic underlying idea of *inverse search* is that (potential) providers of information do not have to reveal or capture their information beforehand, but can use data about actual information needs to evaluate demand [19].

We thus conceptualize inverse search as information providers, matching their information against a given set of information needs - in opposite to conventional search, where information seekers match their information needs (i.e. queries) against a given set of information (i.e. documents). While users "import" public information into their private space in conventional search, inverse search helps

to move information from the private space to the public space, where it might satisfy the information needs of other users.

When we talk about inverse *semantic search*, we consider SPARQL-like structured queries on structured RDF-like data[4]. SPARQL-like queries will be used to estimate the demand for certain metadata triples in a certain knowledge base. With inverse semantic search, we aim to address how and why metadata moves from private to public spaces. Besides sharing existing metadata, demand information can also be used to signal metadata which is not yet captured at all.

Architecture We will now describe a system architecture that supports the envisioned metadata sharing process.

In order to differentiate between metadata in public and in private spaces and to fulfill requirement R1, the system distinguishes a public metadata space ($MSpace_{Public}$) and a private metadata space for each user (e.g. $MSpace_{Chrissy}$ and $MSpace_{Dave}$), which is not accessible to any other user. Technically, this can be realized either by physical or logical separation. Physical separation means, that the private space is an independent system running on the local machine of a user (e.g. a Semantic Desktop system). Logical separation does not require two separate applications, but can be implemented as a feature in a server-based system – e.g. by offering "private" and "public" sharing options.

Queries to the public space are automatically saved to a public *query log*. Both, the public space and the query logs can be accessed by any user. In order to retain privacy (R1), queries may be anonymous and must not contain information about the querying user. However, if users like to receive automatic notifications when new metadata arrives, they might need to reveal their identity.

As functional modules, our approach requires a *SearchApplication* which allows to query both the local and the public space and a *SharingEngine*, which periodically compares the local with the public space and the query log. Again, this can be realized either within a web-application or by combining a public web-based space with an application running on a local machine.

Thus, the sharing engine can provide an estimation of how useful it would be to share certain metadata. This helps to satisfy requirement R3, since the user is guided in her decision which metadata is worth sharing. In order to minimize the effort of sharing (requirement R2), several ways are possible to suggest sharing certain metadata to the user. This might either happen by enriching existing interfaces (e.g. by blending metadata with information about its value [17]) or by periodically presenting a ranked list of sought metadata.

Process At runtime, inverse semantic search comprises a number of subsequent steps, which are lined out in the following. As in the example in section 3.1, this process starts with the collection of information needs and its aggregation. Afterwards, information needs are retrieved by potential information providers

[4] We are well aware that there are many other notions of semantic search

and matched against their private metadata. Finally, they can decide to share or create certain metadata, if it matches some demand.

Information need The information need of information seekers drives our knowledge sharing process. As it is the most convenient source of information needs, we will stick to *queries* resp. *query logs* as our main input and do not discuss other possible sources in this paper.

As driven by the use cases in section 3.2, we assume a fragment of common metadata query languages in the scope of this paper. Based on SPARQL, the most common and standardized language, queries in our approach are restricted concerning the free variables they can contain. We assume a single variable in the result set, which can have arbitrary constraints concerning object property values, literal values and property value types. We allow an additional second free variable to help defining type constraints for object properties. The resulting eight query archetypes haven been presented in Table 1. We now describe the internal ("query log") storage format for these queries.

Result type Instances or literals
List<Type> Type constraints for the resulting instance (?x rdf:type t; not applicable to queries for literals)
List<Instance, Property> List of tuples of instances and properties constraining the result (i, p, ?x)
List<Property, Object> List of tuples of properties and instances constraining the result (?x, p, i; not applicable to queries for literals)
List<Property, Literal> List of tuples of properties and literals constraining the result (?x, p, l; not applicable to queries for literals)
List<Property, Type> List of properties and their type constraining the result (?x, p, t; not applicable to queries for literals)
List<Type, Property> List of types and properties constraining the result (t, p, ?x)
Timestamp Timestamp of the query
User Concrete or abstract user id
Number of results Current number of results for the query in the knowledge base (i.e. size of the result set for the most recent query)

Each query archetype listed in Table 1 will be logged in one of the *List* fields of the log. Combinations of constraints will result in several entries. We refer to query instances in this log (Q) by using the variable i. A query q denotes a set of query instances i, which is similar in all fields except of timestamp, user and number of results.

Need aggregation Need aggregation targets the ranking of queries in terms of identifying those queries which information need is only badly satisfied by the underlying public knowledge base. We therefore apply two processing steps to the data in the query log.

First, identical queries are aggregated on a per-user basis to calculate a *personal information need*. For the sake of simplicity, we only consider identical

queries in the scope of this paper (q; see above). Second, the different personal information needs concerning a particular query are aggregated into an *aggregate information need*. We shortly motivate this distinction, before we describe how to calculate these values.

The information need of a user is a primary subject of investigation in information retrieval (IR). The main purpose of IR systems is to help users satisfying their information needs by providing a set of relevant documents. A personal information need can be defined as information which a user requires to complete a specific task [20]. To use an IR system, the user typically has to express this information need in terms of the query language which can be interpreted by the search system. In most systems, this is a textual, "keyword-based" representation.

Based on this definition of personal information need, we conceptualize *aggregate information need* (AIN) as an aggregate of the personal information needs of members in a group. By *group* we mean the group of users which are able to access a certain public space. Depending on the concrete setup, this can be a team, an organization or the web as a whole. The aggregate information need thus denotes the overall amount of information which the members of this group require to complete their particular tasks.

Our basic rationale for computing the AIN is that it is higher, 1) the more often and more recently a term has been part of a query, 2) the more different users used the term in a query and 3) the more seldom a term is in the local space of the users. Based on this, we define the AIN as the weighted sum of individual information needs of the querying users.

We propose the following four measures as signals for an aggregate information need:

Frequency We assume that the AIN regarding a query is the higher, the more often it has been executed.

Availability If few results are returned for a query, the availability of metadata is low which indicates a higher demand. For availability, the number of results for a query is normalized into an interval $[0, 1]$.

Freshness Since the AIN regarding some query is a dynamic value, we also assume that the AIN is higher, the more recently the query has been executed. This allows recent information needs to score a relatively higher value.

Universality We define that an AIN is the higher, the more different users issued the same query. The rationale behind this is that an information provider may only receive a limited set of sharing recommendations (see section 3.2). Thus, in order to maximize the overall benefit for the organization, such metadata should be prioritized, which is relevant for a large number of different information seekers.

In order to formally define the AIN, we group the first three signals into a personal information need. Thus, the personal information need for a user concerning a specific query consists of the availability, frequency and freshness

of queries:

$$PIN(q, user) = (1 - r) \cdot \frac{Q_{q,user}}{Q} \cdot (1 + \frac{Q_{q,user-recent}}{Q_{recent}}) \quad (1)$$

Accordingly, the AIN is the sum of the values for PIN, normalized by the total amount of querying users:

$$AIN(q) = \frac{Users_q}{Users} \cdot \sum_{user_q} PIN(q, user) \quad (2)$$

Further need aggregation could be done by aggregating structured queries using similarity measures leveraging taxonomic knowledge from a background ontology. This kind of aggregation could potentially be done at server- or client-side. However, further considerations in this direction are out of the scope of this paper.

Need retrieval The "query log" as presented before needs to be available for retrieval by interested metadata providers. Therefore, we define two major services:

List<InformationNeed> getTopInformationNeeds() returns the most desired information needs from the metadata repository under consideration.

List<InformationNeed> getInformationNeedsRelatedTo(URI) returns the information needs w.r.t. a certain instance URI.

Both services return a list of *InformationNeed* objects, which basically represent entries from the query log.

Local matching In the local matching step, the retrieved information need is matched against the private metadata of the information provider. Therefore, the *InformationNeed* objects are transformed back into SPARQL queries, where all attributes are marked as optional. The results are finally ranked by the number of constraints they fulfill. Again, more sophisticated matching approaches are possible, but their discussion is beyond the scope of this paper.

Sharing In terms of the actual user interaction for sharing, several possibilities exist. One option could be to embed the described knowledge sharing mechanism in an existing application (e.g. some kind of knowledge browser). This browser occassionally triggers the information need backend and matches it against the private metadata of the user. Once metadata is identified to be worth sharing, the user interface indicates this e.g. by highlighting the respective data. A concrete example could be a Semantic Wikipedia [21] browser, which identifies sought metadata for a browsed page.

The second option would be to provide an explicit sharing mechanism which presents the user a raw list of sought metadata. The user could then decide to generate this list (e.g. once a week or each time a certain program starts) and share respective metadata. A Semantic Wikipedia example could here be a list of desired metadata within the overall Wiki, similar to the existing list of "Wanted pages" in the MediaWiki software (Special:WantedPages).

4 Conclusion

In this paper, we addressed the issue of why and how metadata is provided for the public Semantic Web. In particular, we introduced a mechanism called *inverse semantic search* which targets to support *knowledge providers*. It is based on the principle of aggregating unsatisfied information needs in order to recommend the sharing or capturing of information. By considering requirements rooted in studies on knowledge sharing (c.f. section 3.2), our system design explicitly considers user incentives [7, 16, 6].

Since a concrete evaluation of this system would be a challenge of its own, it was not in the scope of this paper and is left to future research. However, since related research has shown that meta-information can foster user contributions [17, 16], we are confident that our approach will have practical value. Evaluation would require to incorporate design choices based on different motivational factors into the user interface which allows to test according hypothesises at system runtime (similar to [16]).

Regarding the level of granularity, our discussion was based on the vision of the Semantic Web as such. However, we think that our approach can also be beneficial in more restricted settings such as organizations or teams. Furthermore, the described mechanism could also be built into applications such as Semantic Wikis to guide and foster metadata generation.

Finally, this paper focused on describing the general motivation, architecture and design principles of inverse semantic search. Several technical issues such as the modelling of information needs based on more complex structured queries or the semantic aggregation of queries should be addressed by future work.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic Web. Scientific American **284**(5) (May 2001) 34–43
2. Thomas, C.F., Griffin, L.S.: Who will create the metadata for the internet? First Monday **3**(12) (1998)
3. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM (2007) 971–980
4. Ankolekar, A., Krötzsch, M., Tran, T., Vrandecic, D.: The two cultures: mashing up web 2.0 and the semantic web. In: WWW '07: Proceedings of the 16th international conference on World Wide Web, New York, NY, USA, ACM (2007) 825–834
5. Bizer, C., Cyganiak, R.: D2r server-publishing relational databases on the semantic web (poster). In: International Semantic Web Conference. (2006)
6. Kustanowitz, J., Shneiderman, B.: Motivating annotation for personal digital photo libraries: Lowering barriers while raising incentives. Technical Report HCIL-2004-18, University of Maryland, College Park, MD, USA (01 2005)
7. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Ht06, tagging paper, taxonomy, flickr, academic article, to read. In: HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia, New York, NY, USA, ACM (2006) 31–40

8. Happel, H.J., Stojanovic, L.: Analyzing organizational information gaps. In: Proceedings of the 8th Int. Conference on Knowledge Management. (2008) 28–36
9. Cress, U., Hesse, F.W.: Knowledge sharing in groups: experimental findings of how to overcome a social dilemma. In: ICLS '04: Proceedings of the 6th international conference on Learning sciences, International Society of the Learning Sciences (2004) 150–157
10. Cabrera, A., Cabrera, E.F.: Knowledge-sharing dilemmas. Organization Studies **23** (2002) 687–710
11. Wasko, M.M., Faraj, S.: Why should i share? examining social capital and knowledge contribution in electronic networks of practice. MIS Quarterly **29**(1) (2005) 35–57
12. Ardichvili, A., Page, V., Wentling, T.: Motivation and barriers to participation in virtual knowledge-sharing communities of practice. Journal of Knowledge Management **7**(1) (2003) 64–77
13. Desouza, K.C.: Barriers to effective use of knowledge management systems in software engineering. Commun. ACM **46**(1) (2003) 99–101
14. Desouza, K.C., Evaristo, J.R.: Managing knowledge in distributed projects. Commun. ACM **47**(4) (2004) 87–91
15. Orlikowski, W.J.: Learning from notes: organizational issues in groupware implementation. In: CSCW '92: Proceedings of the 1992 ACM conference on Computer-supported cooperative work, New York, NY, ACM Press (1992) 362–369
16. Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., Kraut, R.E.: Using social psychology to motivate contributions to online communities. In: CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work, New York, NY, USA, ACM (2004) 212–221
17. Rashid, A.M., Ling, K., Tassone, R.D., Resnick, P., Kraut, R., Riedl, J.: Motivating participation by displaying the value of contribution. In: CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, New York, NY, USA, ACM (2006) 955–958
18. Prud'Hommeaux, E., Seaborne, A.: SPARQL query language for RDF. World Wide Web Consortium, Recommendation REC-rdf-sparql-query-20080115 (January 2008)
19. Happel, H.J.: Closing information gaps with inverse search. In: Practical Aspects of Knowledge Management, 7th International Conference. Lecture Notes in Computer Science, Springer (2008) 74–85
20. Baeza-Yates, R., Riberio-Neto, B.: Modern Information Retrieval. ACM Press (1999)
21. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic wikipedia. In: WWW '06: Proceedings of the 15th international conference on World Wide Web, New York, NY, USA, ACM (2006) 585–594

Tackling the Curse of Prepayment – Collaborative Knowledge Formalization Beyond Lightweight

Valentin Zacharias[1], and Simone Braun[1]
[1] FZI, Research Center for Information Science, Haid-und-Neu Strasse 10-14, 76131 Karlsruhe, Germany
{zach, braun}@fzi.de

Abstract. This paper argues for collaborative incremental augmentation of text retrieval as an approach that can be used to immediately show the benefits of relatively heavyweight knowledge formalization in the context of Web 2.0 style collaborative knowledge formalization. Such an approach helps to overcome the "Curse of Prepayment"; i.e. the hitherto necessary very large initial investment in formalization tasks before any benefit of Semantic Web technologies is visible. Some initial ideas about the architecture of such a system are presented and it is placed within the overall emerging trend of "people powered search".

Keywords: Semantic Web, collaborative knowledge formalization, web 2.0, Semantic Wikis, People Powered Search

1 Introduction

The Curse of Prepayment is the Chicken-Egg problem of Semantic-Technologies: that Semantic technologies promise great functionality only after a large amount of knowledge is formalized. And that no one is willing to invest large amounts of money or time in formalization until the great functionality is visible or at least foreseeable.

Recently there has been a great interest in approaches that attempt to tackle this problem by adapting Web 2.0 ideas to make knowledge formalization collaborative, and very easy, cheap, and simple (e.g. [1,2]). In this way these approaches enable end users to successfully contribute to the creation of semantic structures. However, most of these approaches are restricted to very lightweight formalisms – there seems to be a lack of ideas how to extend these approaches to more powerful formalisms. This paper argues that the critical point that stops these approaches from adequately addressing heavyweight formalisms is – again - the Curse of Prepayment: that with these approaches an investment in (more) heavyweight formalization shows no immediate benefit. For example it is trivially possible to edit an OWL Full document in any Wiki by just uploading its XML representation, but there is nothing enabled by the continued development of this document; nowhere is it visible what kind of functionality is made possible by this formalization.

We present the "Collaborative Incremental Augmentation of Text Retrieval" as one approach that can be used to tackle this challenge. It stipulates to enable endusers to

collaboratively and incrementally extend a conventional search engine in the direction of question answering. In section 2 this paper starts with an examination of current approaches in this area and their attempts to tackle the Curse of Prepayment; the chicken-egg problem of Semantic technologies. In section 3, the five properties of simple, collaborative, incremental, partial and immediate are presented as critical in this respect. Section 4 then details the challenges of extending this kind of knowledge formalization to more heavyweight formalisms. Collaborative, incremental augmentation of text retrieval is introduced as one possible answer to this challenge in section 5; some ideas on its realization are contained in section 6. Finally the paper concludes with a short summary and a discussion of connections to related work.

2 Web 2.0 Knowledge Formalization and The Curse of Prepayment

The Curse of Prepayment is also often referred to as the Chicken-Egg problem of Semantic Web technologies: Semantic Web technologies promise great functionality once a large amount of knowledge is formalized. However, because knowledge formalization is difficult, often not well supported, and cumbersome, the investment beforehand needed to see any functionality is very large (cf. [3]). This is problematic, because users cannot learn from seeing the final effects of their changes, are not motivated from seeing growing functionality, and because organization may hesitate to make investments in new technologies when any visible success is very far off.

This is not a new observation and numerous approaches have emerged to address it – of particular interest here are approaches that try to harness Web 2.0 ideas for this task[1]**.** The assumption of these systems can be summarized as "*Maybe formalization can be made so simple and useful and distributed over so many people that people will do it for free*". These approaches can be roughly separated into three groups:

- **Social Semantic Tagging Systems:** Based on the observation that a large number of people are successfully creating structured data with tagging applications, these approaches try to extend these systems with a bit more structure, a bit more formality. Our own SOBOLEO[2] system [4], GroupMe [5], Int.ere.st [6], BibSonomy[3] [7], Fuzzzy[4] [8] and gnizr[5] are examples for these kinds of systems.
- **Semantic Wikis:** The second group of systems starts from the observation that people are spending large amounts of time creating semi-structured data in wikis. These system then try to give people the tools and the support such that they can create data with more structure, more formality. The Semantic Media Wiki[6] [9],

[1]Not mentioned here, but also important are research threads based on machine learning (automatically acquiring structure) and exposing pre-existing structure (e.g. exposing relational databases as SPARQL endpoints)

[2] http://www.soboleo.com

[3] http://www.bibsonomy.org

[4] http://www.fuzzzy.com

[5] http://gnizr.googlecode.com/

[6] http://semantic-mediawiki.org/

Freebase[7], IkeWiki [10] and MyOntology [2] are example for these kinds of systems.

- **Semantic Games with a Purpose:** The third, much smaller, group is inspired by the success of the gwap platform[8], based on the "Games with a Purpose" paradigm [11]. This platform offers games that – as a side effect – also create structured data for the computer. OntoGame[9] is the approach that realized this for the Semantic Web [12]. This approach stands very much apart from the other approaches because (from a user point of view) the goal of the formalization is the formalization itself. This very interesting approach will nevertheless always only be able to address a small subset of needs for formalization and will not be discussed further in this paper.

In the authors' view there are five closely related properties that give these Social Semantic Tagging and Semantic Wikis a chance to tackle the curse of prepayment:

- **Simple:** Formalization is simple, can be done with little training, little effort and not only by logic experts. For example compared to an traditional ontology engineering tool the SOBOLEO and the Semantic Media Wiki are very easy to use.
- **Collaborative:** Formalization can be done jointly in a group – in this way the cost is spread over multiple persons; the prepayment needed from every person is reduced. All Web 2.0 knowledge formalization approaches have collaboration at their core.
- **Incremental:** Not everything needs to be formalized at once, formalization can be done incrementally. With the Semantic Media Wiki system the user can introduce typed relations incrementally as time is available.
- **Partial:** The tools can work with data stores that are only partly formalized, that contain data at different levels of formality. Again in Semantic Media Wiki, for example, typed relations can co-exist with internal links.
- **Immediate:** Formalized data can be used immediately, immediately brings some benefit to the user. With SOBOLEO or BibSonomy the user has an immediate advantage from adding just one 'broader' relation between tags, because his sped up.

Together these five properties can be summarized as: "*Making Every Penny Count, Immediately*". There is an immediate benefit for formalizing even small parts; and because these systems are simple and collaborative, formalizing these small parts is relatively cheap.

Hence in the authors' opinion **this immediate benefit for formalizing even small parts lies at the core of these systems' success.** The exact nature of this benefit differs between systems, examples are:

- **Tables and less redundant data:** The unique selling point of the Semantic Media Wiki: as soon as just a few attribute values have been specified, these can be used to create tables and overview pages that before had to be maintained manually.

[7] http://www.freebase.com/
[8] http://www.gwap.com/gwap/
[9] http://www.ontogame.org/

- **Hierarchical Organization:** In systems like SOBOLEO or BibSonomy tags can be organized hierarchically, this allows for more effective maintenance of the tag repository as well as for more effective navigation and retrieval. This works after having just one such relation.
- **Advanced Search:** For example in the SOBOLEO system adding just one synonym for a tag/concept will already improve the search experience, searching for this synonym will then also consider the documents annotated with the topic.

The immediate benefit is very important because it enables users to learn about the effects of their changes, it can motivate volunteer contributors to continue and finally it can also provide the justification for a continued investment of an organization.

3 The Challenges of Heavyweight Formalization

However, all the 'immediate benefits' presented in the previous section are benefits from very lightweight formalizations:

- **Tables and less redundant data:** The automatically generated overview tables envisioned for Semantic Wikipedia [9] only depend on simple RDF triples.
- **Hierarchical Organization:** The hierarchical organization in BibSonomy depends on just one taxonomic relation without a formal semantic.
- **Advanced Search:** The semantic search of the SOBOLEO system depends only on taxonomic broader-narrower relations and labels.

None of the mentioned systems can show a comparable immediate benefit from e.g. adding rules, disjunction statements, or elaborate models with many different relations between entities. Further, the most powerful of these, the arbitrary queries supported by Semantic Media Wiki can only be used by users with relatively advanced knowledge about the data model and the query language.

Extending the mentioned systems in the direction of more heavyweight formalisms faces many challenges, such as (partially based on [13]):

- **Usability / Debuggability:** Formalisms such as OWL or First Order Logic are harder to understand, in particular faults are much harder to identify.
- **Robustness:** A single faulty statement added to a knowledge base with a millions of axioms can make the knowledge base inconsistent and thereby invalidate all conclusions. Unless this problem is tackled, open collaborative knowledge formalization is impossible.
- **Performance and the Language Expressivity / Performance Tradeoff:** Current reasoners for OWL Full or FOL could not support a continuously updated knowledge base of even a fraction of the size of Wikipedia; hence restrictions on language expressivity, not-sound or incomplete algorithms or some use of non-declarative languages would be needed.
- **Mixed Formality:** Incremental and partial formalization also means that the data store is never fully formalized; always contains data at different levels of formality. Again a challenge for current reasoning approaches.

In the opinion of the authors, however, all of these challenges are trumped by the Curse of Prepayment – the question about the immediate benefit of formalizing even small parts of a data store. What is to be gained from spending some time and/or

money from bringing a part of a data store to a highly formal level, how is this immediately visible to the editors? Knowing an answer to this question may then also allow to find answers to the tradeoffs implied by the challenges above, e.g. this may provide the justification to remove certain powerful but slow features from the knowledge representation language or help decide whether to keep soundness or completeness of the reasoning algorithms used (in cases where both cannot be achieved).

An answer to the Curse of Prepayment for more heavyweight formalism must provide a way to profit from these formalizations that is useful, understandable and immediately visible to the user. This answer needs to realize the five properties of simple, collaborative, incremental, partial, and immediate for heavyweight formalisms.

One way to utilize heavyweight formalism is the creation of question answering systems, i.e. systems that do not just point a user to a document but that rather provide direct answers to questions. However, so far it has been impossible to create question answering systems that can answer the majority of arbitrary user questions, leading to almost constant disappointment of users. A further problem is that the creation of question answering systems for even small domains is a very costly and time consuming process. Also by now users are used to keyword based queries and there is evidence that they prefer keyword based queries to full question answering [14].

The proposed approach stipulates the collaborative creation of a question answering system by incrementally extending a text retrieval system. In this way the question answering functionality can harness the highly formal knowledge, the information retrieval engine prevents disappointment of the users, and the collaboration distributes the cost down.

4 Collaborative, Incremental Augmentation of Text Retrieval

Collaborative, incremental augmentation of text retrieval means the stepwise extension of normal text retrieval in the direction of questions answering. One for one, frequent queries that users already pose to a system are identified and the data store is extended to allow the computation of direct answers to these questions. For examples the maintainers of a site notice that queries of the form “<country name> size” are often entered. They then extend the search engine to detect this pattern and add formalizations needed to directly answer it.

The stepwise augmentation of text retrieval is already visible in modern search engines. For example posing the query “weather Karlsruhe” to Yahoo returns not just pages containing this string but an actual weather report for the city of Karlsruhe. Searching with Microsoft Search and the query “5 EUR in yen” returns the amount of Yen that 5EUR can buy with this days exchange rate. Google even allows developers to extend its search via the subscribed links feature[10]. For example, users subscribed to a Wikimedia Data[11] search extension that pose the query “distance from Paris to Karlsruhe” get the correct result of 443km; a result created through a specific file that

[10] http://www.google.com/coop/subscribedlinks/

[11] http://www.google.com/coop/profile?user=01659747360823524154O

contains the locations of cities based on Wikipedia entries. Yahoo also allows for the extension of its search engine in a related way through the SearchMonkey[12] platform.

Shown above is another example of augmentation of test retrieval – here from the ask.com search engine in response to the query "china size".

This stepwise augmentation of text retrieval in the direction of question answering has a number of advantages:

- **Reasonable Expectations:** No current question answering technique can answer the majority of arbitrary formulated natural language queries. For this reason current question answering systems will answer most queries incorrectly – something very few users are willing to accept. With augmented text retrieval question answering is an added bonus that appears only in relatively well understood cases. It thereby avoids the trap of constantly disappointing the user's expectation.
- **Incremental and Partial:** Functionality to answer queries can be added step by step, possibly depending on the progression of the overall formalization of the data store. No large up-front investment is needed.
- **Immediate:** As soon as the functionality to answer one kind of queries is complete, it can become part of the search engine and improve the user experience.
- **Accepted Interface:** That the system builds on what is currently probably the most accepted interface for information search.

These advantages mirror many of the desired properties identified in the previous sections. What is missing from these systems, however, is the notion of simple and collaborative participation in the creation of these answers. Google's Subscribed Links and Yahoo's SearchMonkey do this to a certain extent, but only for developers that are willing to learn the respective protocols and formats.

We hence propose the collaborative incremental augmentation of text retrieval as the next target for collaborative (Web2.0 style) knowledge formalization approaches. We propose to show the immediate benefit of higher levels of formality by enabling users to incrementally extend an information retrieval engine into a question answering system.

5 Realization

This section details some initial ideas on the architecture and layout of such a system in order to further explain the notion of collaborative incremental augmentation of text retrieval. The section starts with an overview of the question answering process followed by thoughts on the core reasoning architecture.

[12] http://developer.yahoo.com/searchmonkey/

Query processing starts with the user entering a query, as an example the user might enter “china size”. In order for the system to be able to process a set queries of the form “<country name> size” in a common way, it must first detect that some part of the query refers to a country. For this detection step the system uses the data already entered into the system by the users, i.e. the names and synonyms of countries. For the example query “china size” the output of this step might consist of the following:

```
china, fzi/col#Peoples_Republik_of_China
       a physicalThing, a country, a state ...
```

This indicates that the “china” part of the query could be matched to an instance with the URI “fzi/col#Peoples_Republik_of_China that is known to be of types physicalThing, country and state.

In the next step the system matches the processed query against a number of ’templates’ collaboratively maintained in a wiki-like system. These templates specify the relation between queries and internal queries. One such template might be:

```
<#a type:physicalThing> size
=>
        <-    #a a physicalThing
              #a size ?b
```

This defines that a query consisting of the reference to an entity of type “physicalThing” followed by the string “size” is translated into a query of the form shown above. This query mainly looks for a triple of the form #a size ?b, where #a is the country from the query and ?b is the variable representing the searched value. Obviously graphical editors would be needed to support the user in the creation of these templates.

In a next step the query created in this way is processed by the system using all information available. The result of this query processing is then presented together with the result from a normal information retrieval system. Additional (user maintained) templates might support the presentation of results.

The actual processing of the query can be done using any kind of formalization, such as OWL ontologies, FOL axioms, rules or even specialized heuristics created in procedural programming languages. We think that the best approach is not one based on a monolithic knowledge base using only one general purpose reasoned, but rather one build from relatively large heterogeneous reasoning modules; some using DL reasoners, some executing procedural scripts and some using parameterized heuristics. The important aspect is, however, that the elements used by these reasoning modules are created collaborative by the users and that these reasoning modules in their use in the augmented text retrieval then show the benefit of having these highly formal elements immediately.

In this way the proposed system can iteratively grow from an information retrieval system into a question answering system that can use all kinds of heavyweight knowledge for query processing. E.g. the example query introduced above could be processed using mapping rules that mediate between different vocabularies; or it could profit from OWL based reasoning that lead to the inference that a particular entity is a physical thing.

6 Related Work

The presented idea is part of the broader trend of 'People Powered Search'[13]; a trend that tries to unify the search paradigm exemplified by Google with the open, social collaboration of delicious[14] and Wikipedia[15]. Examples for other approaches within this trend are Mahalo[16] and Wikia Search[17] that understand result pages as akin to wiki pages that can be edited. Further examples are 50matches[18] that only searches pages bookmarked in social bookmarking services and sproose[19] that allows voting for results.

Question Answering systems and natural language interfaces have been developed for more than 30 years [15,16], with recent years seeing again a rise in interest in these systems (e.g. [17,18,19,20]); this recent rise fueled by the availability of a plethora of lexical resources, upper level ontologies, of the shelf grammars and parsers and advances in databases and knowledge representation [17]. With AskJeeves the recent years even saw a (now aborted) attempt to bring question answering to mainstream web search. Our proposed approach differs from this strand of research in the following ways:

- **Collaboration:** That the functionality of the system is created during use by its users (and not before)
- **Incremental:** That functionality to answer some queries directly is added step by step. This is only possible because an information retrieval engine forms the backup.
- **Existing Queries:** That users are not encouraged to 'speak to the machine'; that rather queries done anyway are detected.

7 Conclusion

This paper has presented collaborative, incremental augmentation of text retrieval as one answer to the question of what can be the benefit for formalizing parts of a data

[13] Also known as ‚Human Powered Search' or ‚User Powered Search'
[14] http://www.delicious.com/
[15] http://www.wikipedia.org
[16] http://www.mahalo.com/
[17] http://search.wikia.com
[18] http://www.50matches.com/
[19] http://www.sproose.com/

store with more than very lightweight formalisms. In this sense this idea goes beyond existing Web 2.0 style collaborative knowledge formalization approaches that obtain all their direct benefit only from very lightweight formalizations.

The advantages of this approach are (1) that a question answering system is build incrementally, without raising unreasonable expectations (2) that an improvement can be shown almost immediately, after only a small initial investment and (3) that it builds on what is currently probably the most accepted interface for information search.

As the obvious next step we plan to implement this idea as an extension of the SOBOLEO system. This is part of our ongoing project to support all stages of our proposed Ontology Maturing process model for collaborative knowledge formalization.

References

1. Braun, S., Schmidt, A., Walter, A., Nagypal, G., Zacharias, V.: The Ontology Maturing Approach to Collaborative and Work-Integrated Ontology Development: Evaluation Results and Future Directions. In: Proc. of the ESOE-Workshop at ISWC'07, CEUR Workshop Proc. Vol. 292 (2007) 5-18
2. Siorpaes, K., Hepp, M.: myontology: The marriage of ontology engineering and collective intelligence. In: Bridging the Gap between Semantic Web and Web 2.0 (SemNet 07). (2007) 127–138
3. Friedland, N., Allen P., Matthews, G., Witbrock, M., Baxter, D., Curtis, J., Shepard, B., Miraglia, J., Angele, J., Staab, S., Mönch, E., Oppermann, H., Wenke, D., Porter, B., Barker, K., Fan, J., Chaw, S.Y., Yeh, P., Tecuci, D., Clark, P.: Project Halo: Towards a digital Aristotle. AI Magazine, 29(4) (2004) 29-48
4. Zacharias, V., Braun, S. (2007). SOBOLEO – Social Bookmarking and Lightweight Ontology Engineering. In: Proc. of the Workshop on Social and Collaborative Construction of Structured Knowledge at WWW'07, CEUR Workshop Pro. Vol. 273 (2007)
5. Abel, F., Henze, F.M., Krause, D., Plappert & D. Siehndel, P.: Group Me! Where Semantic Web meets Web 2.0. In: Proc. of the 6th Int. Semantic Web Conf. (2007)
6. Kim, H.L., Yang, S.-K., Song, S.-J., Breslin, J.G. & Kim, H.-G.: Tag Mediated Society with SCOT Ontology. In: Proc. of the 5th Semantic Web Challenge at ISWC'07 (2007)
7. Hotho, A., Jäschke, R., Schmitz, C.; Stumme, G.: BibSonomy: A Social Bookmark and Publication Sharing System. In: CS-TIW'06. Aalborg: Aalborg University Press (2006)
8. Lachica, R., Karabeg, D.: Metadata creation in socio-semantic tagging systems: Towards holistic knowledge creation and interchange. In: Scaling Topic Maps. Topic Maps Research and Applications 2007, Springer, (2007)
9. Krötzsch, M., Vrandecic, D., Völkel, M., Haller, H., Studer, R.: Semantic Wikipedia. In: Journal of Web Semantics 5/2007, pp. 251–261, Elsevier (2007)
10. Schaffert, S.: IkeWiki: A Semantic Wiki for Collaborative Knowledge Management. In: 1st International Workshop on Semantic Technologies in Collaborative Applications (STICA 06), Manchester, UK, June (2006)
11. Von Ahn, L.: "Games with a Purpose," Computer, vol. 29, no. 6, 2006, pp. 92–94.
12. Siorpaes, K., Hepp, M.: Games with a Purpose for the Semantic Web. Intelligent Systems, IEEE, (23):50-60,2008

13. Krötzsch, M., Schaffert, S., Vrandecic, D.: Reasoning in Semantic Wikis. In: Proc. of the 3rd Reasoning Web Summer School, Dresden, Germany, vol. 4636 of LNCS, pp. 310--329. Springer (2007)
14. Reichert, M., Linckels, S., Meinel, C., Engel, T.: Student's perception of a semantic search engine, In IADIS Cognition and Exploratory Learning in Digital Age, Porto, Portugal, oo. 139-147 (2005)
15. Ogden, W., Bernick, P.: Using natural language interfaces. In Helander, M., editor, Handbook of Human-Computer Interaction. Elsevier (1996)
16. Copestake, A., Jones, K. S.: Natural language interfaces to databases. Knowledge Engineering Review, 5(4):225–249. Special Issue on the Applications of Natural Language Processing Techniques. 1989
17. Cimiano, P., Haase, P., Heizmann, J., Mantel, M.: Orakel: A portable natural language interface to knowledge bases. Technical report, Institute AIFB, University of Karlsruhe (2007)
18. Kaufmann, E., Bernstein, A., Fischer, L.: Nlp-reduce: A "naıve" but domainindependent natural language interface for querying ontologies. In: 4th ESWC, Innsbruck, A (2007)
19. Bernstein, A., Kaufmann, E., Kaiser, C.: Querying the semantic web with ginseng: A guided input natural language search engine. In: 15th Workshop on Information Technologies and Systems, Las Vegas, NV, pp. 112–126 (2005)
20. Frank, A., Krieger, H.-U., Xu, F., Uszkoreit, H., Crysmann, B., Jörg, B., and Schäfer, U.: Question answering from structured knowledge sources. Journal of Applied Logic, Special Issue on Questions and Answers: Theoretical and Applied Perspectives, 5(1):20–48 (2007)

The 7th International Semantic Web Conference

October 26 – 30, 2008

Congress Center, Karlsruhe, Germany