Crowdsourcing and the Semantic Web (Dagstuhl Seminar 14282)

Bernstein, Abraham ; Leimeister, Jan Marco ; Noy, Natasha ; Sarasua, Cristina ; Simperl, Elena

Abstract: Semantic technologies provide flexible and scalable solutions to master and make sense of an increasingly vast and complex data landscape. However, while this potential has been acknowledged for various application scenarios and domains, and a number of success stories exist, it is equally clear that the development and deployment of semantic technologies will always remain reliant of human input and intervention. This is due to the very nature of some of the tasks associated with the semantic data management life cycle, which are famous for their knowledge-intensive and/or context-specific character; examples range from conceptual modeling in almost any flavor, to labeling resources (in different languages), describing their content in terms of ontological terms, or recognizing similar concepts and entities. For this reason, the Semantic Web community has always looked into applying the latest theories, methods and tools from CSCW (Computer Supported Cooperative Work), participatory design, Web 2.0, social computing, and, more recently crowdsourcing to find ways to engage with users and encourage their involvement in the execution of technical tasks. Existing approaches include the usage of wikis as semantic content authoring environments, leveraging folksonomies to create formal ontologies, but also human computation approaches such as games with a purpose or micro-tasks. This document provides a summary of the Dagstuhl Seminar 14282: Crowdsourcing and the Semantic Web, which in July 2014 brought together researchers of the emerging scientific community at the intersection of crowdsourcing and Semantic Web technologies. We collect the position statements written by the participants of seminar, which played a central role in the discussions about the evolution of our research field.

DOI: https://doi.org/10.4230/DagRep.4.7.25

Report from Dagstuhl Seminar 14282

Crowdsourcing and the Semantic Web

Edited by

Abraham Bernstein[1], Jan Marco Leimeister[2], Natasha Noy[3], Cristina Sarasua[4], and Elena Simperl[5]

1 **Universität Zürich, CH,** bernstein@ifi.uzh.ch
2 **Universität Kassel, DE & Universität St. Gallen, CH,** leimeister@uni-kassel.de
3 **Google Inc. – Mountain View, US,** natashafn@acm.org
4 **Universität Koblenz-Landau, DE,** csarasua@uni-koblenz.de
5 **University of Southampton, GB,** E.Simperl@soton.ac.uk

Abstract

Semantic technologies provide flexible and scalable solutions to master and make sense of an increasingly vast and complex data landscape. However, while this potential has been acknowledged for various application scenarios and domains, and a number of success stories exist, it is equally clear that the development and deployment of semantic technologies will always remain reliant of human input and intervention. This is due to the very nature of some of the tasks associated with the semantic data management life cycle, which are famous for their knowledge-intensive and/or context-specific character; examples range from conceptual modeling in almost any flavor, to labeling resources (in different languages), describing their content in terms of ontological terms, or recognizing similar concepts and entities. For this reason, the Semantic Web community has always looked into applying the latest theories, methods and tools from CSCW (Computer Supported Cooperative Work), participatory design, Web 2.0, social computing, and, more recently crowdsourcing to find ways to engage with users and encourage their involvement in the execution of technical tasks. Existing approaches include the usage of wikis as semantic content authoring environments, leveraging folksonomies to create formal ontologies, but also human computation approaches such as games with a purpose or micro-tasks.

This document provides a summary of the *Dagstuhl Seminar 14282: Crowdsourcing and the Semantic Web*, which in July 2014 brought together researchers of the emerging scientific community at the intersection of crowdsourcing and Semantic Web technologies. We collect the position statements written by the participants of seminar, which played a central role in the discussions about the evolution of our research field.

Seminar July 6–9, 2014 – http://www.dagstuhl.de/14282
1998 ACM Subject Classification I.2.9 Robotics Artificial Intelligence / Robotics, D.3.1 Formal Definitions and Theory – Semantics, H.1.2 User/Machine Systems – Human information processing
Keywords and phrases Crowdsourcing, Human Computation, Games with a Purpose, Microtask Crowdsourcing, Semantic Web, Linked Data, Quality Assurance, Crowd Management, Workflow Management, Interfaces, Gamification, Incentives
Digital Object Identifier 10.4230/DagRep.4.7.25
Edited in cooperation with Cristina Sarasua

Crowdsourcing and the Semantic Web, *Dagstuhl Reports*, Vol. 4, Issue 7, pp. 25–51
Editors: Abraham Bernstein, Jan Marco Leimeister, Natasha Noy, Cristina Sarasua, and Elena Simperl

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Abraham Bernstein
Jan Marco Leimeister
Natasha Noy
Cristina Sarasua
Elena Simperl

The aim of the *Dagstuhl Seminar 14282: Crowdsourcing and the Semantic Web*, which was held in July 2014, was to gain a better understanding of the dual relationship between crowdsourcing and Semantic Web technologies, map out an emerging research space, and identify the fundamental research challenges that will need to be addressed to ensure the future development of the field.

The seminar focused on three categories of topics: first and foremost we looked into existing crowdsourcing approaches and how these could or have been applied to solve traditional semantic data management tasks. Particular attention was paid to core components of a crowdsourcing-enabled data management and processing system, including methods for quality assurance and spam detection, resources, task and workflow management, as well as interfaces, and the way these components can be assembled into coherent frameworks. A second category of topics that was addressed during the seminar reached out to other disciplines such as economics, social sciences, and design, with the aim to understand how theories and techniques from these fields could be used to build better crowdsourcing-enabled data management systems for the Semantic Web. Last, but not least, we discussed the usage of semantic technologies within generic crowdsourcing scenarios, most notably as means to describe data, resources and specific components.

The seminar, in its community-formative role, represented the starting point for the emergence of working groups that will in the future jointly address the identified scientific challenges. Participants were asked to provide a 1-page position statement reflecting on why they think it makes sense to consider the two topics – crowdsourcing and Semantic Web (or Web of Data) – at the same seminar. Specifically, participants were asked to write a statement reflecting on one of both of the following questions:

1. What are the Semantic Web tasks where you felt you needed crowdsourcing? Why? What were the challenges?
2. What are the crowdsourcing tasks where using semantics might help? Why? What are the challenges?

The first two days of the seminar were dedicated to presentations of topics related to position statements and working groups on use case scenarios and challenges identified during the talks and Q&A sessions. The third day focused on the consolidation of the results of the working groups and the definition of next steps and follow-up activities.

In the following sections we present the position papers written by the researchers of the crowdsourcing and the Semantic Web community, who took part in the seminar. We will publish a more complete research roadmap for crowdsourcing and the Semantic Web at a later stage.

2 Table of Contents

Executive Summary

Abraham Bernstein, Jan Marco Leimeister, Natasha Noy, Cristina Sarasua, and Elena Simperl . . . 26

Position Papers

Crowdsourcing Linked Data Management
Maribel Acosta . . . 29

Using Crowdsourcing for Semantic Annotation in Media Sector
Sofia Angeletou . . . 30

Semantic Interpretation and Crowd Truth
Lora Aroyo . . . 31

Check and Track – Crowdsourcing Semantics and Semantic Crowdsourcing
Irene Celino . . . 31

Crowdsourcing & the Semantic Web
Philippe Cudré-Mauroux . . . 32

Crowdwork and Microtasks
Roberta Cuel . . . 33

Crowdsourcing is for the Tail
Gianluca Demartini . . . 34

Crowdsourced Feature Selection
Michael Feldman . . . 35

Crowdsourcing Semantic Tasks for Scientific Research
Yolanda Gil . . . 36

Position Statement Dagstuhl Seminar on Crowdsourcing and the Semantic Web
Carole Goble and Steve Pettifer . . . 37

Crowdsourcing Platforms and the Semantic Web
Atsuyuki Morishima . . . 38

Crowdsourcing and the Semantic Web one-page Position Statement
Valentina Presutti . . . 39

Crowsourcing Ontology Lexicalization
Philipp Cimiano . . . 40

Towards Hybrid-genre and Embedded Crowdsourcing
Marta Sabou . . . 41

Crowdsourcing for Evaluation and Semantic Annotation
Harald Sack . . . 43

Who and How Should Be Involved in Crowdsourced Data Interlinking?
Cristina Sarasua . . . 44

Linking Implicit with Explicit Semantics: An Initial Position Statement
Markus Strohmaier . . . 46

Opinions and Aims in Participatory Sensing
Gerd Stumme . . . 46

Crowdsourcing and the Semantic Web
Tania Tudorache 47

The Role of Crowdsourcing and Semantic Web for Consumable APIs
Maja Vukovic 48

Training Systems with Crowd Truth
Christopher A. Welty 49

Merging Contexts
Marco Zamarian 50

Participants 51

3 Position Papers

This section includes the complete list of position statements, which participants (except for the organizers) provided.

3.1 Crowdsourcing Linked Data Management

Maribel Acosta (KIT – Karlsruher Institut für Technologie, DE)

The Semantic Web is envisioned as a system in which machines can truly understand the meaning of requests posed by humans or other machines, assuming that all the data consumed and produced in this system is encoded with semantics. For this vision to become a reality, it is necessary in the first place to create semantically enriched data that will allow machines to have such comprehension of the data. Furthermore, even in the presence of semantic data, there are still data enhancement tasks to be addressed that require the execution of processes that are intrinsically better performed by humans than by machines, like disambiguation, association, pattern recognition, etc. In particular, considering the limitations of machines when the meaning of the data is highly contextual or subjective, we can foresee a powerful use of crowdsourcing approaches for Linked Data management tasks for the following problems:

Link prediction

Involves the creation of links on the instance level. Link prediction includes the problem of entity resolution by creating owl:sameAs links as well as the creation of other type of links between resources. This is particularly challenging when the data lacks specificity (no useful additional information is provided), has many homonyms (requiring a disambiguation step), and its variety is high. Therefore, the decision to create a link or not is highly influenced on the context and type of the data.

Data quality assessment

Refers to the process of validating data to detect inconsistencies or other type of errors in the data. In order to successfully execute this task, it is necessary to discern the possible types of errors encountered in the data to perform the appropriate corrective actions.

In the previously discussed tasks, human intervention can serve different purposes. For instance, crowdsourcing approaches could perform all the single steps of a task. The main challenge in this case is to ensure scalability in terms of monetary cost and execution time when executing very large tasks. Another strategy would be to apply crowdsourcing to generate training data, e. g., ground truth datasets, and implement supervised learning techniques. While this approach is more flexible in terms of scalability, since learning approaches are highly sensitive to training data, this strategy requires the creation of high quality data from the crowd. As a complement, crowdsourcing techniques can also be applied to validate intermediary outcomes of automated approaches.

3.2 Using Crowdsourcing for Semantic Annotation in Media Sector

Sofia Angeletou (BBC – London, GB)

One of the challenges in the uptake of semantic technologies and linked data is the lack of high quality, semantically annotated content that can be reused and repurposed using linked data principles. Although there is a plethora of open vocabularies available, the volume of good quality annotated content across various domains is not comparable. Such a gap hinders the creation of applications that could make use of such content and showcase the value of this technology in real world use cases; either in the public domain (open data) or internally in organisations.

In this paper I argue that using crowdsourcing is a means to obtaining high quality semantically annotated content with a special focus on the Media sector. The primary output of organisations in this sector includes the publication of content assets such as programmes, news articles and educational works both in a linear broadcasting but also on an "on demand" basis. The cases for using semantically annotated content vary from improved management of media assets, creation of curated content indexes with a strong semantic layer and novel audience facing applications that involve personalised content offerings based on the things that matter to each individual member of the audience.

Current observations in some Media organisations show that semantic annotations are inconsistent both in completeness and in quality, rendering the quality of the consuming product poor. The semantic tagging techniques may involve entity extraction as a first step, but then rely on a single editor to apply or approve relevant tags. This does not always yield a good tagging result, given the subjectivity of the tagger and various workflow factors, such as availability of time available for annotation. Having the manual annotations weighted based on the popularity of selected concepts can contribute to the solution of the problem to a large extent. Crowdsourcing could be encouraged either internally in an organisation or opened up to the public, to allow interested volunteers contribute to the annotations. However, both cases pose interesting challenges.

In internal crowdsourcing, the process should be designed such that it fits in the workflow of content editors without creating additional overhead, given their restricted time. In addition they should be informed of the impact of their action as a motivation to continue contributing. In many cases the value of tagging is not clear to the content editors or annotators, it is often seen as an additional and disconnected task they must complete that causes delays in the accomplishment of other production activities.

In cases where crowdsourcing could be employed in an open world setting there are a few challenges that would influence both the result but also the perception of organisations about implementing such practices. Some forms of crowdsourcing would typically involve technically competent users of a service, who might be biased towards particular selections not representative of the whole of the audience of an organisation. In addition, allowing the public to annotate content would require further editorial control to ensure that the annotation results reflect the editorial policies and branding of the organisation.

3.3 Semantic Interpretation and Crowd Truth

Lora Aroyo (Free University of Amsterdam, NL)

Human annotation is a critical part of semantic processing, from the early days of knowledge acquisition to modern methods of collecting data to train and evaluate machine learning algorithms that perform semantic interpretation tasks. However, conventional human annotation, and indeed Semantic Web technologies themselves, are based on an antiquated ideal of a single correct truth that needs to be disrupted. I propose a new theory of truth, Crowd Truth, that is based on the intuition that human interpretation is subjective, and that measuring annotations on the same objects of interpretation (e. g. sentences, images, videos) across a crowd will provide a useful representation of their subjectivity and the range of reasonable interpretations. I will introduce several metrics for measuring quality of human annotated data based on Crowd Truth, and present experimental results that show these metrics are inter-related in a way that previous human annotation methodologies did not reveal.

3.4 Check and Track – Crowdsourcing Semantics and Semantic Crowdsourcing

Irene Celino (CEFRIEL – Milano, IT)

The Semantic Web needs Crowdsourcing for a number of different tasks; however, in my opinion, the most fruitful role for Crowdsourcing is for data quality and validation in the Semantic Web. The vast popularity and success of the Linked Data initiative has been bringing an increasingly large amount of data on the Semantic Web or Web of Data, often in form of raw data or automatically-generated/transformed information. Thus, the issue of checking data quality, correctness, consistency and update becomes of utmost importance [7, 6, 3] and here is where I see a concrete use of Crowdsourcing techniques. In my experience of building Human Computation games for geo-spatial data management [2], the results showed that data curation (especially in case of outdated information) was definitely a plus with respect to pure data collection, in terms of precision and outcome value. In the context of Semantic Web research, Crowdsourcing has been employed also for other knowledge-intensive tasks (e. g. ontology engineering or ontology alignment [5]), at times also successfully. Still, I believe that the simplicity of the crowdsourced task should be always taken into due consideration, thus data curation could be a better purpose for Crowdsourcing rather than ontology management; in any case, the issue of selecting the right crowd is always open. Moreover, fact checking is definitely a domain in which human judgement easily outperforms automatic techniques, so it is probably a more relevant objective for applying Crowdsourcing techniques.

Crowdsourcing needs the Semantic Web for crowd users' tracking: the issue of understanding who the crowd workers are, their background and expertise, and ultimately their reliability is central to any Crowdsourcing effort. This is why I believe that Semantic Web technologies can be key to describe people profiles as well as to track the provenance of an information

value chain (i. e., workflow of the data lifecycle). Indeed, models derived or inspired by W3C PROV-O have being adopted to keep trace of human interventions [1] or to log modifications to a triple dataset [4] (collected both via crowdsourcing and heuristic/statistical approaches). As a consequence, Semantic Web techniques can be employed to supervise the Crowdsourcing process, in terms of evidence collection, agreement and decision making, and especially crowd users' tracking and evaluation. Tracing the crowdsourced information can be very useful to compare different strategies to aggregate results as well as to provide incentives and rewards to the crowd.

References

1 I. Celino. Human Computation VGI Provenance: Semantic Web-based Representation and Publishing. IEEE Transactions on Geoscience and Remote Sensing, 51(11):5137–5144, 2013.
2 I. Celino. Geospatial Dataset Curation through a Location-based Game. Semantic Web Journal, 5(7), 2014.
3 I. Celino, E. Della Valle, and R. Gualandris. On the effectiveness of a Mobile Puzzle Game UI to Crowdsource Linked Data Management tasks. In 1st International Workshop on User Interfaces for Crowdsourcing and Human Computation, 2014.
4 M. Knuth and H. Sack. Data Cleansing Consolidation with PatchR. In Posters and Demos – ESWC 2014, 2014.
5 C. Sarasua, E. Simperl, and N. Noy. CrowdMap: Crowdsourcing Ontology Alignment with Microtasks. In 11th International Semantic Web Conference, pages 525–541, 2012.
6 E. Simperl, B. Norton, and D. Vrandecic. Crowdsourcing Tasks within Linked Data Management. In Proceedings of COLD2011, volume 782. CEUR-WS.org, 2011.
7 J. Waitelonis, N. Ludwig, M. Knuth, and H. Sack. Whoknows? evaluating linked data heuristics with a quiz that cleans up dbpedia. Interactive Technology and Smart Education, 8(4):236–248, 2011.

3.5 Crowdsourcing & the Semantic Web

Philippe Cudré-Mauroux (University of Fribourg, CH)

The Semantic Web was, from its inception, destined to create a machine-processable Web of data, a Web where computerized agents could collect, integrate, exchange and reason upon large quantities of heterogeneous online information. Over the years, however, new applications of the Semantic Web emerged. Today, some of its most exciting applications—such as the display of semi-structured information related to an entity, or the parsing of natural language for text summarization or question answering—are directly targeting human users. In that sense, the Web of data is increasingly used to help humans in their daily lives. In contrast, crowdsourcing leverages a Web of documents, made for humans, to help machines solve computationally complex tasks. Crowdsourcing and the Semantic Web were in that sense bound to meet each other, and to come together to solve some of the most formidable open problems in information management. In my research group, the eXascale Infolab[1] both topics frequently support each other. On one hand, we recently used crowdsourcing to help solve complex Semantic Web issues such as entity linking [1] or instance matching [2]. Semantic

[1] http://exascale.info/

data, on the other hand, was instrumental in facilitating advances in push crowdsourcing [3] and in crowdsourced data sensing [4]. Another interesting convergence might occur in a few years, when generalized micro-task crowdsourcing platforms will emerge and host arbitrary complex tasks—some annotated using Semantic Web information. Both human and machines might then compete for the same tasks on the crowdsourcing infrastructure, creating de facto and for the first time a universal and hybrid Semantic Web service infrastructure for information processing.

References

1 G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In WWW, pages 469–478, 2012.

2 G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Large-scale linked data integration using probabilistic reasoning and crowdsourcing. VLDB J., 22(5):665–687, 2013.

3 D. E. Difallah, G. Demartini, and P. Cudré-Mauroux. Pick-a-crowd: tell me what you like, and i'll tell you what to do. In WWW, pages 367–374, 2013.

4 M. Wisniewski, G. Demartini, A. Malatras, and P. Cudré-Mauroux. Noizcrowd: A crowd-based data gathering and management system for noise level data. In MobiWIS, pages 172–186, 2013.

3.6 Crowdwork and Microtasks

Roberta Cuel (University of Trento, IT)

Organizations increasingly use crowdsourcing to solve problems outside the reach of traditional work processes. By having access to a practically unlimited pool of contributors providing an unprecedented spectrum of skills, experiences and ideas, organizations are enabled to organize their knowledge, increase the flexibility of their processes, and drive openinnovation.

Crowdsourcing for Semantic Web

Organizations can use crowdsourcing to organize their unstructured information and knowledge, identifying entities and disambiguating meaning of words, images, videos, etc. This semantic knowledge can later be used to improve searching and reasoning processes in the corporate knowledge memories, knowledge bases, archives, etc. This development comes with a multitude of social, legal, technical and economic challenges. In this page I want to focus on two of these. How people understand and find an agreement upon a piece of information structure. Especially in big corporates, workers belonging to different units (R&D, production, marketing, accounting, etc.) use different interpretation schemas to identify entities and disambiguate meaning. Indeed, depending on different interpretation schemas, people may use the same categories with different meanings or different words to mean the same thing. An entity can be considered as the "explicit part of what we know" and gets its meaning from a typically implicit taken for granted interpretation schemas (among others see paradigms [Kuhn, 1979], frames [Goffman, 1974], thought worlds [Dougherty, 1992], context [Ghidini & Giunchiglia, 2001], mental spaces [Fauconnier, 1985], cognitive path [Weick, 1979], etc.) How workers deal with crowdsourcing and their daily activities. People's participation and willingness to contribute are critical issues that organizations

must take into account when introducing crowdsourcing solutions. In the specific case of a corporate setting, we are in a principal agent relationship in which the two parties (employer and workers) have different interests and asymmetric information. In order to ensure high quality results and reduce moral hazard and conflict of interests a set of incentive should be designed.

Semantic Web for Crowdsourcing

Semantic web can enable crowdsourcers (content providers) to provide a better quality contributions. In the particular case of microtasks, such as document/image annotations and very simple semantic disambiguation, crowdsourcers can take advantage from the suggestions provided by a semantic based system. We developed a semantic based platform for the disambiguation of Diagnosis-related groups (DRG) in an hospital setting. Doctors have improved their contributions, getting suggestions about what DRGs better suite the description of a patient record. In that case, Semantic Web helped doctors to choose a concept among a pre-elaborated set of entities, reducing time of patient record elaboration and DRG recognition, improving the precision of the choices and the variability in the DRG recognition. One of the big challenges is the process of knowledge convergences or divergences that may emerge do to the suggetions provided by the semantic based system. When people have to choose among a selection of pre-defined items, tend to use them as they are, and the final result is often a common and unique conceptualization of knowledge, not very innovative.

References

1 E. Von Hippel, Horizontal Innovation Networks U by and for Users, working paper 4366-02, MIT Sloan School of Management, 2002.
2 R. Cuel et al., Motivation Mechanisms for Participation in Human-Driven Semantic Content Creation, Int'l J. Knowledge Eng. and Data Mining, vol. 1, no. 4, 2011, pp. 331–349.
3 N. Kaufmann, T. Schulze, and D. Veit, More than Fun and Money, Worker Motivation in Crowdsourcing U A Study on Mechanical Turk, Proc. 17th Americas Conf. Information Systems, AIS, vol. 1, no. 11, 2011, pp. 1–11.
4 C. Prendergast, The Provision of Incentives in Firms, J. Economic Literature, vol. 37, no. 1, 1999, pp. 7–63.
5 D. B. Brabham, Moving the Crowd at Threadless: Motivations for Participation in a Crowdsourcing Application, Information, Communication, and Society, vol. 13, no. 8, 2010, pp. 1122–1145.

3.7 Crowdsourcing is for the Tail

Gianluca Demartini (University of Fribourg, CH)

Semantic Web needs crowdsourcing for data quality issues. However, the seminar should just briefly overview the standard and obvious tasks like: Entity linking (attach URIs to entities in text), knowledge base integration (identify the same entity over two KBs and create owl:sameAs statements), ontology matching (identify the same predicates in two ontologies), relation creation/check (generating or validating RDF predicates), and knowledge base quality curation (fact checking). For sure, within this set of tasks an open question is how

to optimally design HITs for the crowd trying, for example, to avoid Semantic Web specific terminology like ontology, predicate, URI, etc.

The aspect I would like to discuss the most is to, which, data crowdsourcing should be applied in order to make hybrid human-machine approaches suitable at Web scale. Human computation approaches (either crowdsourcing or editorial curation) are already successfully applied to very popular entities: Examples include the Google Knowledge graph and Wikipedia where either employees or crowds manage the content and improve its quality. The open challenge remains for tail, entities, that is, the very many different entities which are not popular or valuable enough individually but would have a great value as a whole (e. g., collecting all small restaurants opening hours in a city).

The point I want to make at the seminar is that micro-task Crowdsourcing has to be used for tail entities and not just for entities worth to appear in Wikipedia or in the Google Knowledge Graph. In order to make the crowd effectively work on those entities we need to leverage worker skills and passions [1] or the communities they belong to [2]. The question then becomes about which crowd should be used for a specific task. To better answer this question, that is, to find the right workers in the crowd for a task, we need to leverage profiling techniques, recommender, systems, and push, crowdsourcing as we are currently doing with OpenTurk [3]. In other cases when the right worker is not available, it is necessary to train the crowd before it can perform some tasks, as, for example, dealing with domain specific (e. g., medical) entities.

References

1 Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Pick-A-Crowd:Tell Me What You Like, and I'll Tell You What to Do. In: 22nd International Conference on World Wide Web (WWW 2013), Rio de Janeiro, Brazil, May 2013.

2 Michele Catasta, Alberto Tonon, Djellel Eddine Difallah, Gianluca Demartini, Karl Aberer, and Philippe Cudré-Mauroux. Hippocampus: Answering Memory Queries using Transactive Search. In: 23rd International Conference on World Wide Web (WWW 2014), Web Science Track. Seoul, South Korea, April 2014.

3 http://alpha.openturk.com

3.8 Crowdsourced Feature Selection

Michael Feldman (Universität Zürich, CH)

I am currently interested in feature selection for classification tasks based on wisdom of crowd. So far this domain was a prerogative of machine learning algorithms, responsible for extracting most substantial features such that by knowing them the entity will be classified with minimal error [3]. However, while this approach is effective for explicit and structured data, extracting features to classify non-trivial, logical structures is extremely challenging for traditional machine learning algorithms [2], [1]. For instance, to classify writer by her writing style may be done by natural language processing methods as Latent Semantic Analysis. However, these methods have limited success with significant drawbacks as scalability, synonymy or polysemy treating. Therefore I hope to explore during the workshop the following aspects:

1. Challenging Semantic Web tasks that can not be resolved efficiently with existing tools, but the performance of such tools may be boosted by extracting tacit knowledge of crowds.

2. Exploring the existing methods of crowd engagement to solve typical Semantic Web problems.
3. Possible ways to explore the tacit knowledge of crowds and to outline it in succinct way (e. g., as a set of features or significant patterns) with regards to relevant Semantic Web tasks. All this, taking in account existing Semantic Web tools and ability to verify the extracted data contribution by boosting performance of these tools.
4. To gain an understanding of how to conceptualise Semantic Web tasks to general framework. As tasks may differ in their definitions (e. g., writing style vs. data integration), I would like to explore approaches providing the means to define general problem while parameters of the tasks are different.

References

1 Alelyani, S., Tang, J., Liu, H.: Feature selection for clustering: A review. Data Clustering: Algorithms and Applications p. 29 (2013).
2 Liu, H., Motoda, H., Setiono, R., Zhao, Z.: Feature selection: An ever evolving frontier in data mining. Journal of Machine Learning Research-Proceedings Track 10, 4–13 (2010).
3 Song, Q., Ni, J., Wang, G.: A fast clustering-based feature subset selection algorithm for high-dimensional data. Knowledge and Data Engineering, IEEE Transactions on 25(1), 1–14 (2013).

3.9 Crowdsourcing Semantic Tasks for Scientific Research

Yolanda Gil (Information Sciences Institute, University of Southern California, US)

We are investigating the use of semantic wikis to develop platforms that enable the flexible incorporation of contributors into science tasks. Our research in this area is in three major fronts:

1. **Organic Data Curation:** There are many datasets in geosciences that have been collected over the years. Even if the datasets are available in centralized repositories, they rarely have metadata. Organizing and describing these datasets are great candidate tasks for crowdsourcing. We are investigating how to form communities that can create metadata for repositories that contain thousands of scientific datasets. We are developing an organic data curation framework that extends semantic wikis so that the contributors can dynamically define new metadata attributes on the fly, and to use those attributes to describe the properties of datasets. We are incorporating in our framework the incentives and rewards that work best for scientific researchers and citizen volunteers to contribute to these tasks.
2. **Organic Data Science:** According to some estimates, 85% of geosciences data is often kept by individual investigators and not available in shared repositories, often called "darked data." Many global environmental science research cannot be carried out without accessing this data. We are developing an organic data science framework where the task of sharing and curating data is an integral part of doing science, and where data contributors can participate in tasks that involve the use of their data.
3. **Tracking the Use of Semantic Wikis in Science:** Semantic wikis have been extensively used to collect and structure scientific knowledge. We are analyzing the content and growth of these wikis as platforms for doing science using Semantic Web technologies. We

have developed ProvenanceBee, a service that collects semantic content and provenance of hundreds of semantic wiki sites. The semantic wiki platform is used very differently in different scientific applications. Our research focuses on studying semantic wiki communities to understand how contributors add semantic content to wikis.

3.10 Position Statement Dagstuhl Seminar on Crowdsourcing and the Semantic Web

Carole Goble and Steve Pettifer (University of Manchester, GB)

What are the Semantic Web tasks where you felt you needed crowdsourcing?

Making scientific knowledge in the literature accessible to machines by stealth.

Why? Currently the knowledge is captured in databases and the scientific literature; these are growing at a rate that makes it impossible for humans to keep up with latest info without support from computers to find and summarize important discoveries.

Databases have at least some structure and are intended to be machine-readable, so there is scope for converting these into formats that are friendlier to the Semantic Web. The literature on the other hand is largely designed for humans; it is full of ambiguities, subtle nuances and rhetorical flourishes that work for human readers but confuse machines. This, and the fact that much of the literature is stored in formats that are hard to process by machine, makes extracting semantics from articles very hard.

What were the challenges

- Much important information is captured in natural language, diagrams or tables. These are hard for a computer to process.
- Legacy formats such as the PDF are typically un-semantic bags of words and lines/curves are difficult to process reliably into anything structured.
- The hedged language used in scientific writing to avoid over-claiming makes identifying claims hard (i. e. we don't say "A does B to C", but "We hypothesize that it may be the case that A, under certain circumstances might behave in a B-like way in the context of C").

What are the crowdsourcing tasks where using semantics might help?

Paywalls and licenses make bulk-mining of the literature by an individual hard/impossible/illegal, but nothing prevents a single person mining a single paper to which they have legitimate access. Doing this en-masse and combining the results would yield a valuable body of machine processable knowledge.

Why? However, text and data mining algorithms are heuristic; semantics are necessary to normalize / cross-validate results to gain confidence, and to manually correct/curate errors in the automatic extraction.

What are the challenges? It is difficult to persuade busy scientists to do work that is for the collective good without any immediate payback; therefore micro tasks have to be built

into a system where the user is either a 'side effect' of something they want to do anyway, or is low-effort and has obvious immediate benefit to them.

Project Lazarus

The purpose of Project Lazarus [1] is to attempt to crowd-source information from the scientific literature 'by stealth', as a side effect of providing scientists with a tool that makes navigating and exploring the scientific literature easier and more rewarding than by conventional means. Scientists thus gain an immediate benefit, and as a side effect of using the tools contribute to a central repository of semantically rich and openly licenced knowledge.

Utopia Documents [2] is a PDF reader that provides many immediately useful features; for example the ability to dereference and 'click through' references to retrieve the cited article without needing to search online, or to extract data from tables into spreadsheets. Project Lazarus aims to extend the tool to capture the 'exhaust gasses' of such activities. With a user's permission, the data from these actions can automatically be contributed to a central repository creating (for example) a citation network, or a repository of data-from-tables. The system can automatically associate provenance with the data, i. e. which user did the extraction, using which algorithm, when, and from which scientific article.

References

1 BBSRC BB/L005298/1 The Lazarus Project: Resurrecting data and knowledge from life science articles by crowd-sourcing, http://www.bbsrc.ac.uk/pa/grants/AwardDetails.aspx?FundingReference=BB/L005298/1.
2 http://getutopia.com

3.11 Crowdsourcing Platforms and the Semantic Web

Atsuyuki Morishima (University of Tsukuba, JP)

Crowdsourcing is a promising tool to solve some of the problems in the semantic Web domain. We are operating a crowdsourcing platform named Crowd4U with the help of researchers from more than 22 universities and conducting several projects for the academic and public purposes. In this paper, we first explain Crowd4U and a project running on it, named L-Crowd, in which we try to clean bibliographic data by crowdsourcing performing tasks for entity identification in bibliographic records. Then, we discuss lessons learned that might be related to the Semantic Web.

Crowd4U

Crowd4U is a microtask-based crowdsourcing platform which is similar to Amazon mechanical turk. It is unique in several ways compared to similar systems. First, Crowd4U provides a high-level abstraction for complex human-machine computation. Second, it supports various task assignment and incentive structures including push/pull-style task assignments. Because many workers voluntarily perform tasks on Crowd4U, they are called *contributors*. Many of these contributors are university students. They performed more than 1,100 tasks per day

in May 2014. The estimated number of anonymous workers on Crowd4U is more than one thousand.

L-Crowd

Crowd4U is hosting several non-commercial crowdsourcing projects.

L-Crowd is one of these projects that started by LIS and CS researchers in Japan to apply crowdsourcing technologies to library problems. In 2012, they designed microtasks to identify different books that have the same ISBN in an effort to clean the bibliography database of the National Diet Library

Lessons Learned and Challenges

From our experience of operating a crowdsourcing platform and conducting several crowdsourcing projects, we think the followings are important related to the Semantic Web.

(1) What are the Semantic Web tasks where we felt we needed crowdsourcing? Entity identification is definitely an important task that needs crowdsourcing. It requires workers to understand the context and conduct some inference based on the background knowledge. Another important task might be to connect concepts to each other in different languages.

(2) What are crowdsourcing tasks where using semantic might help? First, semantic can help workers understand the instructions in the task. Second, semantic can show workers possible results for the task so that workers can choose one of them instead of thinking up their own answer. Third, semantic can improve to data quality of the task results, because it can identify unlikely results for the tasks.

From our point of view, an important challenge is to identify fundermental functions that crowsourcing platforms should provide to support Semantic Web applications and to use semantic to support effective crowdsourcing.

3.12 Crowdsourcing and the Semantic Web one-page Position Statement

Valentina Presutti (CNR – Rome, IT)

What are the Semantic Web tasks where you feel you need crowdsourcing?

- Building cognitive-based resources: resources such as WordNet, FrameNet, etc. should be built and/or validated by means of crowdsourcing in order to reflect the way humans cognitively organize and use their knowledge;
- User-based evaluations: for evaluating methods supporting cognitiveoriented tasks, e. g. content/entity summarization, exploratory search (serendipity, discovery), knowledge relevance, etc., as well as more technical tasks, e. g., formal representation of natural language, ontology learning, property and concept alignment, etc.;

Why?

- because statistical significance requires good numbers of users involved;
- because cognition is a human feature;
- because we need to understand human cognitive capability;

- because it would save time;
- because we need to cope with subjective (context-based) perspectives;
- because I think golden standards-based evaluations often do not apply to Semantic Web research.

What are the challenges?

- identifying the right crowd;
- will to commit to perform high-quality work;
- ensuring high-quality standards (e. g. extremely important for cognitivebased resources);
- reducing complex tasks to simple ones: can we conceive a sort of "reduction" method (cf. complexity theory) to formally or at least rigorously define simplified versions of SW tasks to be assigned to the crowd;
- developing methods that ease modeling crowdsourcing tasks;
- finding incentives for experts to commit to crowdsourcing;

3.13 Crowsourcing Ontology Lexicalization

Philipp Cimiano (Bielefeld University, DE)

For all those applications in which natural language is used to access and interface Semantic Web data, knowledge is needed about how the vocabulary used to describe the data is expressed in natural language. The current state of affairs is one in which each application needs to derive the relation between natural language and formal vocabulary anew. To avoid this situation, lexicon models such as the Lexicon Model for Ontologies (lemon) [3] or the Ontology Lexicon Model (ontolex) have been and are been developed, allowing one to represent this knowledge in a declarative fashion, thus supporting the sharing of this knowledge across applications. Nevertheless, such lexicon models need to be populated, which is a costly process, as different variants of how to refer to a particular vocabulary element in a particular language need to be included. We refer to the task of specifying all the different (lexical) variants that can be used to refer to a particular vocabulary element in some vocabulary or ontology as ontology lexicalization. As an example, consider the property <http://dbpedia.org/ontology/spouse> in DBpedia. This property can be verbalized as follows in English:

X is married to Y X married Y

X is the spouse of Y X is the wife of Y

X is the husband of Y

X is the better half of Y, etc.

As a proof-of-concept of the lemon lexicon model, a manual lexicon for DBpedia has been created [1]. So far, semi-automatic approaches to induce ontology lexica from a corpus have been proposed [2]. In spite of using a corpus-based approach, the work of Walter et al. has shown that human validation is still needed to reach an appropriate quality of the final resource. This, however, is an issue, as human labour needs to be found and rewarded for their work in some way or another (either through payment or other incentives). Crowdsourcing might play a key role in this, involving workers in both the specification of lexical variants as well as the validation of entries proposed by an automatic approach such as the one by Walter et al. Relevant research questions in this context are the following ones:

1. How might a crowdsouring-based framework for ontology lexicalization look like?
2. How should the tasks be formulated to maximize effectiveness and efficiency?
3. By which heuristics or automatic checks can overall quality be ensured?
4. What skills are needed by workers to accomplish the task?
5. What is the cost of lexical pattern acquisition for each vocabulary element?
6. How can workers be motivated to perform the task other than via (micro-) payments?
7. Would a methodology working for English scale to other languages?

References

1 Christina Unger, John McCrae, Sebastian Walter, Sara Winter, Philipp Cimiano: A lemon lexicon for DBpedia. In: Proc. of the NLP & DBpedia Workshop, collocated with ISWC 2013.

2 Sebastian Walter, Christina Unger, Philipp Cimiano: A Corpus-Based Approach for the Induction of Ontology Lexica. In: Proc. of the 18th Int. Conf. on Applications of Natural Language to Information Sytems (NLDB), 2013.

3 John McCrae, Dennis Spohr, Philipp Cimiano: Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In: Proc. of the Extended Semantic Web Conference (ESWC) 2011.

3.14 Towards Hybrid-genre and Embedded Crowdsourcing

Marta Sabou (MODUL Universität Wien, AT)

We see (1) cross-genre crowdsourcing and (2) a tighter integration of crowd-work into ontology engineering as exciting challenges of applying crowdsourcing in the Semantic Web area. Many open issues relate to semantically describing crowd-worker profiles, crowdsourcing tasks, data and workflows to enable advanced functionalities such as (1) flexible matchmaking between tasks and workers or (2) automatic discovery and combination of complex crowdsourcing workflows.

Crowdsourcing for Semantic Web

The Semantic Web task we aimed to solve by crowdsourcing was creating high quality domain ontologies based on an initial semantic structures extracted automatically by an ontology learning algorithm. We needed multiple crowdsourcing tasks, which can be grouped into two categories:

1. Verifying extracted content: are the extracted concepts correct and relevant for the domain? are the extracted subsumption (and other types of) relations correct?
2. Creating new content that is still challenging to extract automatically: given two concepts, is there a relation between them? and if yes, what is that relation?

To accomplish these tasks, we used both games with a purpose (GWAPs) and paid-for crowdsourcing approaches, concluding that both genres have their pros and cons [1]. GWAPs require a higher upfront investment for designing and building them, but the collected contributions are free. Challenges include: designing appealing games; attracting and maintaining a high number of players. In contrast, mechanised labour tasks take less time and effort to set up, can be outsourced to large pools of workers with highly diverse

qualifications, but each contribution costs. On the down-side, it is challenging (1) to phrase tasks in ways that are understood by layman; (2) to ensure the quality of the collected data (e. g., through task design, gold data etc); (3) to find the optimal task setups that maximizes quality while reducing overall cost and completion times. We highlight the following research challenges in applying crowdsourcing for the Semantic Web:

How Can Multiple Crowdsourcing Genres Be Combined? Given the complementary strengths and weaknesses of the GWAPs and paid for crowdsourcing mechanisms, is it possible to combine them in a beneficiary way? For example, crowdsourcing tasks could be split into workflows of tasks, where simpler (e. g., verification) tasks are crowdsourced for money while challenging (e. g., content creation) tasks are solved through game playing. We proposed hybridgenre crowdsourcing as a potential approach [2], but many open issues remain.

How to Embed Crowd-work Into Ontology Engineering? The use of crowdsourcing in the Semantic Web community has matured enough to move on from isolated approaches towards a methodology of where and how crowdsourcing can efficiently support ontology engineers. We see the derivation of such methodologies as an important prerequisite for popularizing the use of crowdsourcing by ontology engineers. Such methodological guidelines should inform the development of tools that facilitate easily embedding crowdsourcing into ontology engineering workflows (for example, extensions of ontology editors).

Semantic Web for Crowdsourcing

Matchmaking between crowd-workers and tasks is an increasing challenge as crowdsourcing platforms attract larger worker bases with diverse qualifications and crowd- sourcing is used for novel and diverse tasks. Semantically representing worker profiles and task types in order to create an optimal matchmaking between them could be an interesting task to solve using Semantic Web technologies. This would allow migrating from pull- to push-based crowdsourcing mechanisms. A key challenge here lies in determining a meaningful category of tasks and finding large enough platforms where these technologies could be implemented and meaningfully tested.

Semantic Description of Crowdsourcing Data and Workflows. Data obtained through crowdsourcing should be represented together with relevant metadata such as: the method used to derive the data; the details of the crowdcontributors; the aggregation method used etc. Such metadata allows for the correct interpretation and exchange of the crowdsourced data. While some vocabularies have been proposed to represent crowdsourced data (e. g., I. Celino's Human Computation Ontology[2]), convergence towards standard vocabularies for this purpose is an important future step. Not only the data of crowdsourcing processes but also their internal workflows could be represented using semantic models, drawing upon and adapting, for example, on previous work on Semantic Web Services.

References

1 Sabou, M., Bontcheva, K., Scharl, A., and Föls, M. (2013). Games with a Purpose or Mechanised Labour?: A Comparative Study. In Proc. of the 13th International Conference on Knowledge Management and Knowledge Technologies, i-Know'13, pages 1–8. ACM.

[2] http://swa.cefriel.it/ontologies/hc

2 Sabou, M., Scharl, A., and Föls, M. (2013). Crowdsourced Knowledge Acquisition: Towards Hybrid-genre Workflows. International Journal of Semantic Web and Information Systems, 9(3):14–41.

3.15 Crowdsourcing for Evaluation and Semantic Annotation

Harald Sack (Hasso-Plattner-Institut – Potsdam, DE)

Evaluation

In general for several evaluation tasks related with Semantic Web applications crowdsourcing is one viable way to achieve a sufficient number of human evaluators for qualitative evaluation. We have already applied crowdsourcing for the evaluation of the following tasks: Fact Ranking [1], Semantic Search, Exploratory Search [2], and Content-based Recommender Systems. All 4 tasks have in common that there is no unique and generally accepted best solution or gold standard available. The most important fact for a given context, the best recommendation based on an example, the best new direction for exploratory search or semantic search. Most times the question of what is the "best result" lies in the eye of the beholder and is dependent on the personal context of the person you ask. Therefore, to create a generally accepted gold standard or evaluation of achieved results, a large number of evaluators must state their opinion about what is best and/or how to rank achieved results. One of the challenges is of course the selection of a representative sample of tasks to be evaluated by a representative (most desirably unbiased in any sense) sample of evaluators. Moreover, for evaluation it is also very important that the evaluators are not cheating. Therefore, an emphasis has to be put on the detection of any kind of fraud within the crowdsourced evaluation task. Also care has to be taken with the decision how to create the task to be solved by the user. It should not be obvious for the user how to influence the outcome of the overall evaluation task in any sense. On the other hand, the task must not be boring to be able to attract a sufficient number of users. Likewise it is also critical that the users possess sufficient expertise to be able to solve the evaluation task in a meaningful way.

Semantic Annotation

A classical task to be solved via crowdsourcing is the provision of (semantic) annotations of any kind of document. This ranges from annotating text with the most suitable keywords, categories, or semantic entities, to the annotation of multimedia documents, such as e. g. images, audio, or video. This is a classical crowdsourcing scenario when dealing with simple textual annotations. With semantic annotations in the sense that predefined semantic categories or even a larger number of potentially suitable semantic entities must be selected by the user, special care has to be taken when designing the user interface in an efficient way. Annotation should be possible in a simple and comfortable way. One example to solve this issue is the provision of auto-suggestion services [3]. In the same way as autocompletion of as e. g. query terms for search engines, autosuggestion takes into account the current user (text) input and suggests potentially suitable semantic entities that fit best to the user's input. If large knowledge bases are used for this task, as e. g. DBpedia, then also a large number of entities might be suggested due to ambiguities and partial matches. Thus, also for auto-suggestion, it is necessary to rank the suggested entities in a way that the entity which

is most likely to be selected by the user is presented first. This is a problem similar to the previously mentioned fact ranking. For image annotation and esp. for time-based (region-based) video annotation, the design and implementation of an efficient user interface is still an important and not completely solved task. For both tasks, evaluation as well as semantic annotation the design of a game-based approach to collect or to produce the necessary user provided data, seems to be a promising way to attract a large number of users [3, 4]. On the other hand, this approach is also rather expensive in terms of developing time and costs. Thus, it must be decided whether to spend resources for game development or as micro payments, as e. g. in mechanical Turk applications.

References

1 A. Thalhammer, M. Knuth, and H. Sack: Evaluating entity summarization using a game-based ground truth, in The Semantic Web – ISWC 2012 (P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. Parreira, J. Hendler, G. Schreiber, A. Bernstein, and E. Blomqvist, eds.), Lecture Notes in Computer Science, vol. 7650, pp. 350–361, Springer Berlin Heidelberg (2012)

2 J. Waitelonis, H. Sack: Towards exploratory video search using linked data, Multimedia Tools and Applications, Volume 59, Number 2 (2012), 645-672, DOI: 10.1007/s11042-011-0733-1, Springer Netherlands, 2012.

3 J. Osterhoff, J. Waitelonis, H. Sack: Widen the Peepholes! Entity-Based Auto-Suggestion as a rich and yet immediate Starting Point for Exploratory Search, 2. Workshop Interaktion und Visualisierung im Daten-Web (IVDW 2012)

4 J. Waitelonis, N. Ludwig, M. Knuth, H. Sack: Whoknows? – Evaluating Linked Data Heuristics with a Quiz that cleans up DBpedia. International Journal of Interactive Technology and Smart Education (ITSE), Emerald Group, Bingley (UK), Vol. 8, 2011 (3)

5 L. Wolf, M. Knuth, J. Osterhoff, H. Sack: RISQ! Renowned Individuals Semantic Quiz – A Jeopardy like Quiz Game for Ranking Facts, 7th Int. Conference on Semantic Systems, ACM Int. Conf. Proc. Series, ACM Inc, (i-Semantics 2011), Graz (Austria), Sep. 7–9, 2011, pp. 71–78

3.16 Who and How Should Be Involved in Crowdsourced Data Interlinking?

Cristina Sarasua (Universität Koblenz-Landau, DE)

The heterogeneity of independently curated data sets on the Web of Data makes the data interlinking process challenging for automatic techniques. With this position statement I would like to describe the topics that arose while researching how to integrate microtask crowdsourcing in the process of automatic RDF data interlinking, and which may be relevant in other knowledge-intensive tasks. For example, dynamically identifying the knowledge of particular crowd workers, taking decisions accordingly and analysing the different kinds of interaction between crowd workers and automatic techniques.

Semantic Web Task with Crowdsourcing

The Semantic Web task where I used crowdsourcing is data interlinking, the process of discovering links between data of different RDF data sets. I used microtask crowdsourcing

as a mechanism to involve humans in the process because they are still crucial to support purely automatic interlinking technology in 1) identifying the sources to be connected, 2) training active learning interlinking algorithms and 3) post-processing the outcome of the automatic interlinking techniques. Microtask crowdsourcing provides a cost-effective and scalable alternative to having expert users in the process, and automates a task that is otherwise often not accomplished systematically. My research focuses on the application of microtask crowdsourcing to the specific scenarios of ontology alignment (i. e. mappings between vocabulary terms) [2] and instance data interlinking (i. .e. links that show equivalence or any other domain-specific relationship between particular individuals). I encountered several challenges that I would like to discuss: first, I experienced that crowdsourcing a task like data interlinking in domain-specific scenarios (e. g. research data and publications of the social sciences) requires crowd workers to be further instructed both in the task and the domain. A second challenge was to accurately identify how suitable the knowledge of a worker is for processing a particular link. Third, the a priori assessment of the value that the crowd can add to the automatic interlinking of a particular pair of data sets. In my opinion, implementing a mechanism to attract and redirect appropriate crowd workers to available microtasks in online marketplaces could optimise crowdsourced work.

Including Semantics in a Crowdsourcing Task

A crowdsourcing task where I see that semantics might help is the recruitment and selection process in microtask crowdsourcing. A machine-readable and interoperable description of the different crowdsourcing agents (i. e. crowd workers and requesters) and their work experience, can promote the recognition for work across the different microtask crowdsourcing platforms. I recently proposed the use of an ontology-based Crowd Work CV [1], to provide detailed descriptions analogously to traditional CVs. This could indirectly help in quality assurance in for example, crowdsourced interlinking microtasks. Another task of the crowdsourcing process where introducing semantics could obviously be useful is in publishing the results. Some crowdsourcing platforms already enable requesters to publish crowd-generated data. If such data was automatically annotated with existing vocabularies (e. g. with the Ontology for Media Resources) and offered as Linked Open Data, it could be more easily consumed and integrated by third-party applications. The challenges that I identify in these areas are: first, the definition of a common understanding of the Crowd Work CV model, which satisfies the requirements of all microtask platforms and can be easily integrated with other business-related information. Second, the tradeoff between ranking crowd workers based on their CVs and giving them the freedom to work on what they are interested in. Third, the adoption of Semantic Web standards by the crowdsourcing platforms.

References

1 Sarasua, C., and Thimm, M. 2013. Microtask available, send us your cv! In Proceedings of the International Workshop on Crowd Work and Human Computation (CrowdWork 2013), co-located with Social Computing and its Applications (SCA2013).

2 Sarasua, C.; Simperl, E.; and Noy, N. F. 2012. Crowdmap: Crowdsourcing ontology alignment with microtasks. In Proceedings of the 11th International Semantic Web Conference (ISWC2012).

3.17 Linking Implicit with Explicit Semantics: An Initial Position Statement

Markus Strohmaier (GESIS & Universität Koblenz-Landau, DE)

Linking implicit with explicit semantics has been one of the original challenges for the Semantic Web: How can we semantically annotate web pages such that both machines and humans are augmented in exploration, navigation and cognitive tasks such as understanding? Much progress has been made with regard to for example linking data sources (e. g. via Linked Data), annotating HTML pages (via e. g. Schema.org) or annotating scientific literature (Bioannotator). Yet these attempts can only be seen as a first step towards a more comprehensive and more systematically interwoven Semantic Web that seamlessly integrates implicit semantics with explicit semantic representations irrespective of type, format or medium. For example, algorithmically annotating images with adequate and valid semantic descriptors still represents a major challenge. Semantically annotating short texts or the novel language (emoticons, slang, hashtags, etc) that is emerging on social media such as Twitter or Facebook represents another challenge that is far from being solved by the current state of semantic and/or natural language understanding methods and techniques. Finally, assigning accurate semantic descriptors to academic or other domain-specific textual resources that require deep background knowledge for proper understanding represents another example of a challenge that has not yet been met.

At the same time crowdsourcing has emerged as an interesting alternative solution to problems that can not be solved with algorithmic approaches alone. Human judgements organized in micro workflows, and augmented with algorithmic approaches for optimization and quality control have the potential to expand our arsenal of algorithms with a flexible, on-demand oracle that could help address some of the fundamental challenges for the Semantic Web. Understanding and managing the potential of crowdsourcing for linking implicit with explicit semantics should represent a pressing challenge for Semantic Web research and the Semantic Web community at large. This needs to include tackling questions related to the design of proper incentive structures, the development of adequate approaches to quality assurance and to novel approaches to evaluation.

3.18 Opinions and Aims in Participatory Sensing

Gerd Stumme (Universität Kassel, DE)

One of the imminent societal challenges is climate change. Both the avoidance of further climatical changes as well as adaptation to them requires significant changes of our societies and economies and of our individual and collective life styles. The enforcement of novel policies may be triggered by a grassroot approach, with a key contribution from information and communication technology. Nowadays low-cost sensing technologies allow the citizens to directly assess the state of the environment; social networking tools allow effective data and opinion collection and real-time information spreading processes. Moreover theoretical and

modeling tools developed by physicists, computer scientists and sociologists allow to analyse, interpret and visualize complex data sets.

The EveryAware project integrates all crucial phases (environmental monitoring, awareness enhancement, behavioural change) in the management of the environment in a unified framework, by creating a new technological platform combining sensing technologies, networking applications and data-processing tools; the Internet and the existing mobile communication networks will provide the infrastructure hosting such platform, allowing its replication in different times and places. Case studies concerning different numbers of participants will test the scalability of the platform, aiming at involving as many citizens as possible thanks to low cost and high usability. The integration of participatory sensing with the monitoring of subjective opinions is novel and crucial, as it exposes the mechanisms by which the local perception of an environmental issue, corroborated by quantitative data, evolves into socially-shared opinions, and how the latter, eventually, drive behavioural changes.

Enabling this level of transparency critically allows an effective communication of desirable environmental strategies to the general public and to institutional agencies.

3.19 Crowdsourcing and the Semantic Web

Tania Tudorache (Stanford University, US)

What are the Semantic Web tasks where you felt you needed crowdsourcing?

The work I am doing is in the context of the collaborative authoring and maintenance of large biomedical ontologies. There are several tasks in which crowdsourcing could be beneficial: (1) Knowledge acquisition – filling out property values for particular ontology entities (e. g., body part associated to a disease from a predefined value set, or synonyms); (2) label translation (e. g., translating medical titles, definitions, synonyms, etc. to different languages); (3) quality assurance, such as, (a) verifying the class-subclass relations in a disease taxonomy, (b) verifying existing property values, (c) verifying that a textual definition corresponds to a formal definition and vice-versa (e. g., the necessary and sufficient conditions of a class description appear as intended in the textual definition); (4) mapping (e. g., mapping between entities in a disease ontology to entities in another medical ontology).

Why? Even though these tasks require some domain knowledge, they are suitable for crowdsourcing as they can be fairly easily sliced into smaller and independent tasks. Some of the tasks may require additional knowledge from the ontology, but this is usually localized and easily extractable.

Challenges. (1) Finding the workers with the appropriate domain knowledge; (2) For task 1, filling out property values from a value set, the challenge is how to present or even filter the values sets in the task, especially if the value sets are larger or hierarchical; (3) expert curation of the crowdsourcing results – once the crowdsourcing results are back, how do the domain experts "vet" them. This is especially important for the knowledge acquisition task of large ontologies, in which high quality results are expected.

What are the crowdsourcing tasks where using semantics might help?

Crowdsourcing could benefit from access to structured and interlinked data, for example: (1) Creating qualifying questions: well established ontologies could provide the source for creating the qualifying questions (e. g., both the taxonomy and property values could be used to generate the questions). (2) Creating "intelligent" and adaptable tasks in an iterative workflow: ontologies and vocabularies can provide the knowledge for generating iteratively enhanced tasks. The results of a task together with the information from the ontology will be used to generate more specific tasks by filtering or pruning out invalid knowledge (e. g., in a combinatorial problem, excluding a taxonomy branch, if the parent was excluded). (3) Provide context for a task: some tasks require background knowledge, which can be more easily obtained from the Linked Data cloud, as the information is structured, rather than from an unstructured source. (4) Quality assurance: similar to task 1, the information from established ontologies and vocabularies could be used to create trick questions that quality workers are expected to respond in conformance with the ontology.

Why? Information on the Semantic Web and Linked Data cloud provide a structured access to data that makes it easier to use in a programmatic way. SW data also provides well-agreed upon background knowledge in form of ontologies and vocabularies that can be useful for different tasks (see above).

Challenges. (1) Identifying the "right" ontologies and vocabularies to be used for a certain domain and task. (2) Identifying the minimum context for a task that can be extracted from the Linked Data cloud, which provides the most useful information to a worker. (3) Providing an easy to use interface (e. g., in forms of software libraries) for working with Semantic Web data (access to linked data or ontologies) for different parts of the crowdsourcing workflow.

3.20 The Role of Crowdsourcing and Semantic Web for Consumable APIs

Maja Vukovic (IBM TJ Watson Research Center – Yorktown Heights, US)

Loose coupling and scalability characterize RESTful (REST) architectures. Simplicity of REST Application Programming Interfaces (APIs) has resulted in rapid development of highly consumable services, through power of reuse [1, 2]. This opens up a significant opportunity to create new service capabilities based on existing REST APIs. This cocreation results in diffused networks of API providers and consumers.

Most commonly APIs are described in terms of endpoints, data formats, and protocols to ease the consumption or even composition of APIs. Researchers are considering novel semantic models and graph-based methods for API discovery [3]. For example, API graph [4] aims to enrich the API descriptions semantically to facilitate not only consumption, but also the provisioning of APIs and the evolution of the API ecosystem. It provides API consumers with information about valuable API composition. It lets API providers benefit from competitive analysis with other APIs. Finally, it enables ecosystem providers to identify gaps in API capabilities and their demand. Early models about incorporating semantics into API networks have emerged [4]. Moreover, this offers opportunities for automatically testing if the data exposed by the APIs follow the principles of linked data [5] so that they can be

interlinked and become more useful. As a result, building a better semantic understanding of each API, its use, and attributes can minimize some of the consumability risks. And the key role here is the one of expert crowd, developers, consumers, and integrators.

Crowd is commonly taking the role of a sensor, actuator and controller in a variety of computing [6]. My interest is in identifying the best way to enable crowdsourcing to enrich the API semantic descriptions, which will lead to more consumable services. For example, can we engage developers, by providing crowdsourcing tasks to label and validate their APIs (within the IDE)? How can consumers participate in this process more effectively? What is the right "task size" and incentive to provide in this domain?

References

1 Pautasso C., O. Zimmermann, and, F. Leymann. Restful web services vs. big'web services: making the right architectural decision. Proceedings of the 17th international conference on World Wide Web. ACM, 2008.

2 DuVander A., "9,000 APIs: Mobile Gets Serious," ProgrammableWeb, http://blog.programmableweb.com/2013/04/30/9000-apis-mobile-gets-serious/, 2013.

3 Dojchinovski M., J. Kuchar, T. Vitvar, and M. Zaremba. Personalised Graph-based Selection of Web APIs. In Proceedings of the 11th International Semantic Web Conference (ISWC). Lecture Notes in Computer Science, no. 7649. Springer Berlin Heidelberg, 2012.

4 Wittern E., J. Laredo, M. Vukovic, V. Muthusamy, A. Slominski. A Graph-based Data Model for API Ecosystem Insights. IEEE International Conference on Web Services. 2014.

5 Bizer C., T. Heath, and T Berners-Lee. Linked data-the story so far. International Journal on Semantic Web and Information Systems (2009): 1–22.

6 Vukovic, Maja, R. Das, S. Kumara. From sensing to controlling: the state of the art in ubiquitous crowdsourcing. International Journal of Communication Networks and Distributed Systems. Volume 11, Number 1. 2013. Inderscience Publishers.

3.21 Training Systems with Crowd Truth

Christopher A. Welty (IBM TJ Watson Research Center – Hawthorne, US)

Crowdsourcing is often used to gather annotated data for training and evaluating computational systems that perform semantic interpretation, such as natural language processing. Crowd workers are asked to perform the same semantic interpretation as the computational system to establish a "ground truth". I will discuss the use of Lora Aroyo's Crowd Truth paradigm to collect human annotated data from the crowd for training a machine learning component that performs the NLP task of relation extraction, with examples and experimental results. I will argue that our results support the hypothesis that semantic technologies reliance on "single truth" styles of semantics are flawed and need to be revisited.

3.22 Merging Contexts

Marco Zamarian (University of Trento, IT)

It seems obvious that we are still far from producing a reliable, automated, method for generating even relatively simple semantic content in cases where context matters. Here, human input is clearly needed and crowdsourcing technologies can help tap this input. A fairly good example might be the case of the development of a common understanding between separated communities sharing a common problem but not a common language. In these cases the exchange of information is typically flawless within each community but becomes almost impossible between the communities. The problem is exacerbated when the two communities do not share a similar level of understanding (i. e. in terms of context) of the problem, and thus basic vocabularies do not overlap in semantic terms. This is the case of rare diseases in medicine, where often there are two (lets consider the simple case of an established syndrome, and not the more complex case of the process of discovery of a new one) clearly defined communities, the experts conducting research on the syndrome on one side, and the patients on the other. The first community shapes up by a process of mutual, scientific recognition, and develops a highly contextualized vocabulary to exchange information on the disease, creating the conditions to, at least in principle, solve the semantic problem. On the other hand, the second community faces a much more challenging situation. Its members are more isolated; they do not have the basic knowledge to access relevant information (even a simple online search by keywords might be beyond their knowledge); in almost all cases they need an interface (namely their family practice doctors) to get access to the first community; their concerns (and thus the kind of information they might be searching for) are totally different from the ones occupying the community of experts. Links and fruitful exchanges of information in these cases occur when common, shared events are organized, however events of this kind are infrequent. Thus, building a platform allowing these separated communities to share information, on the one hand seems like an almost ideal setting for the application of semantic technologies, however spurring the two communities to engage in the production of contents that can be fruitfully exchanged asks for a thoughtful approach. Crowdsourcing, per se, can be thought of as an exercise in developing complex sets of incentives that spur people to engage in activities that can easily become very complex, to the point of being beyond the reach of each contributor. In this specific case, which one can claim to be a typical example of classes of situations where developing vocabularies spanning different communities or different contexts is the goal, the challenges are many and they can be separated into at least two broad categories. First, it is obvious that we are talking about separable communities of potential crowdworkers: devising ways to filter and separate access is non obvious and it could involve ambiguities (what community does a common physician belong to?). Second, there is the need to create separate incentive schemes to participate for members of the two communities. Obviously the experts would be able to contribute more, and more in-depth to the semantic tasks, but are less keen to do so (creating a shared representation is less useful for them, and sharing could potentially undermine their status in their community). The opposite would be true for non-experts.

Participants

- Maribel Acosta
KIT – Karlsruher Institut für Technologie, DE
- Sofia Angeletou
BBC – London, GB
- Lora Aroyo
Free Univ. of Amsterdam, NL
- Abraham Bernstein
Universität Zürich, CH
- Irene Celino
CEFRIEL – Milano, IT
- Philippe Cudré-Mauroux
University of Fribourg, CH
- Roberta Cuel
University of Trento, IT
- Gianluca Demartini
University of Fribourg, CH
- Michael Feldman
Universität Zürich, CH
- Yolanda Gil
University of Southern California – Marina del Rey, US
- Carole Goble
University of Manchester, GB
- Robert Kern
IBM Deutschland – Böblingen, DE
- Jan Marco Leimeister
Universität Kassel, DE & Universität St. Gallen, CH
- Atsuyuki Morishima
University of Tsukuba, JP
- Natasha Noy
Google Inc. – Mountain View, US
- Valentina Presutti
CNR – Rome, IT
- Marta Sabou
MODUL Universität Wien, AT
- Harald Sack
Hasso-Plattner-Institut – Potsdam, DE
- Cristina Sarasua
Universität Koblenz-Landau, DE
- Elena Simperl
University of Southampton, GB
- Markus Strohmaier
Universität Koblenz-Landau, DE
- Gerd Stumme
Universität Kassel, DE
- Tania Tudorache
Stanford University, US
- Maja Vukovic
IBM TJ Watson Research Center – Yorktown Heights, US
- Christopher A. Welty
IBM TJ Watson Research Center – Hawthorne, US
- Marco Zamarian
University of Trento, IT