Human Computation (2015) 2:1:3-17

ISSN: 2330-8001, DOI: 10.15346/hc.v2i1.2

Crowdsourcing and the Semantic Web: A Research Manifesto

CRISTINA SARASUA, UNIVERSITY OF KOBLENZ-LANDAU

ELENA SIMPERL, UNIVERSITY OF SOUTHAMPTON

NATASHA NOY, GOOGLE INC.

ABRAHAM BERNSTEIN, UNIVERSITY OF ZURICH

JAN MARCO LEIMEISTER, UNIVERSITY OF ST. GALLEN, UNIVERSITY OF KASSEL

ABSTRACT

Our goal with this research manifesto is to define a roadmap to guide the evolution of the new research field that is emerging at the intersection between crowdsourcing and the Semantic Web. We analyze the confluence of these two disciplines by exploring their relationship. First, we focus on how the application of crowdsourcing techniques can enhance the machine-driven execution of Semantic Web tasks. Second, we look at the ways in which machine-processable semantics can benefit the design and management of crowdsourcing projects. As a result, we are able to describe a list of successful or promising scenarios for both perspectives, identify scientific and technological challenges, and compile a set of recommendations to realize these scenarios effectively. This research manifesto is an outcome of the Dagstuhl Seminar 14282: Crowdsourcing and the Semantic Web.

---

1. INTRODUCTION

The Semantic Web was designed as a machine-processable Web of data—a Web where computerized agents could collect, integrate, exchange, and reason upon large quantities of heterogeneous online content (Shadbolt et al., 2006). After more than a decade of research and development, the original idea has largely materialized; the foundations of semantic technologies can be found at the core of many success stories in ICT, from Google's Knowledge Graph (Singhal, 2012) to IBM's Watson (Ferrucci et al., 2010). Semantic Web technologies (Hitzler et al., 2009) have become useful in various application areas including domain modeling, data integration, enhanced search and content management within several activity areas such as cultural heritage, health care, public administration and digital libraries (Baker et al., 2012). However, just as many other technologies aiming at automation in a decentralized global environment, they remain reliant on human input and intervention (Siorpaes and Simperl, 2010; Bernstein, 2012). This is due to several factors, but primarily to the knowledge-intensive and context-specific nature of many Semantic Web tasks. Almost any core aspect in the life cycle of semantic content—from conceptual modeling, describing resources in ontological terms and labeling them in different natural languages, to recognizing related concepts and entities in multiple knowledge bases—requires a certain degree of human involvement. Consequently, the Semantic Web community is exploiting theories, methods, and tools from other disciplines, including online communities, participatory design, crowdsourced human computation and collective intelligence with the ultimate goal of building the "Global Brain Semantic Web" (Bernstein, 2012), a Semantic Web including distributed interleaved human-machine computation.

Crowdsourcing offers a cost-effective way to distribute the completion of a task among a potentially large group of contributors (Howe, 2008; Quinn and Bederson, 2011). No matter which specific approach one follows (i.e., volunteering-based, paid, gamified), it has become a very useful means to realize hybrid information processing architectures in which crowd and machine intelligence are brought together to tackle tasks that computers alone find difficult to solve (Bernstein, 2013). In this document we use the term *'crowdsourcing'* broadly to refer to any of the models which can be applied to achieve such goals, including paid microtask and macrotask crowdsourcing (Kittur et al., 2013), enterprise crowdsourcing (Vukovic and Bartolini, 2010), citizen science (Raddick et al., 2009) and other online communities of volunteers (Blohm et al., 2013), GWAPs (games-with-a-purpose) (von Ahn and Dabbish, 2008), 'stealth' crowdsourcing (also known as 'side-effect computing') (Doan et al., 2011), participatory sensing (Burke et al., 2006), and combinations of these. Particulars aside, *contributors* are members of the crowd who participate in the crowdsourcing process, while *requesters* are people or organizations outsourcing *tasks* to the crowd. Besides directed crowdsourcing, we also consider collaborative and passive crowdsourcing (Bigham et al., 2015).

Our goal with this research manifesto is to define a research roadmap for the emerging field at the intersection between crowdsourcing and the Semantic Web. When analyzing the interplay of semantic and crowdsourcing technologies, we explore two categories of scenarios: (i) those in which Semantic Web tasks are approached by seeking the involvement of the crowd (see Section 2); and (ii) those in which the design and operation of a crowdsourcing process is enhanced through the use of machine-processable semantics (see Section 3).

This research manifesto is an outcome of the *Dagstuhl Seminar 14282: Crowdsourcing and the Semantic Web* (Bernstein et al., 2014), which in July 2014 brought together 26 members of this emerging research community to reflect upon the synergies between these two topics and discuss directions for future research.

2. CROWDSOURCING FOR THE SEMANTIC WEB

We see several areas in which crowdsourcing can be a valuable tool in the Semantic Web endeavour. First, it can enable (at least in some cases) an improvement in accuracy of existing automatic techniques by offering a systematic way to augment these with human inputs. Second, it achieves this in a scalable and affordable way, by distributing tasks to a large number of contributors and using novel incentive models to encourage participation. Third, it allows us to exploit the cognitive diversity of collective intelligence. This vision may be applied to a wide range of application domains (e. g. literature management (Morishima et al., 2012), urban and geospatial systems (Celino et al., 2012b; Atzmueller et al., 2014), the media sector (Raimond et al., 2014) and the medical domain (Mortensen et al., 2013a)).

2.1. Scenarios

In each of the Semantic Web scenarios in which these benefits may be exploited, the scope and purpose of crowd contributions and the gestalt of the human-computer interaction can vary greatly (Simperl et al., 2013b). Sometimes, it is about creating training data that an algorithm uses to improve its outcomes. In other cases, the crowd is asked to intervene in the execution of an algorithm to validate intermediary results or is even involved in the design of the overall workflow and the configuration of its run-time parameters. An example for the latter are hybrid query processing or search engines e.g., (Demartini et al., 2013; Acosta et al., 2015).

An alternative way to look at this space is by considering the kinds of activities the crowd is expected to perform, which some crowdsourcing sources group into data collection, data analysis, and problem-solving (Shadbolt et al., 2013). The first refers to those situations in which new data, in our case new Semantic Web artifacts such as ontologies, knowledge bases, data sets of some sorts, benchmarks, gold standards and so on are created or enriched through crowd contributions. This is different from data analysis scenarios, in which the crowd is asked to examine specific properties of these artifacts; in terms of Semantic Web tasks this would mean questions such as identifying correspondences between concepts and entities or classifying instance data into ontological types. In the third case the crowd is confronted with a challenge to solve according to a set of predefined criteria; the focus there is less on breaking down the task into smaller chunks that are taken on by different contributors, but on designing holistic solutions to a given problem, very much in the spirit of open innovation (Leimeister et al., 2009; Boudreau and Lakhani, 2013). The three classes of crowd activities are by no means disjunct and it is likely that any real-world crowdsourcing exercise will include elements of collection, analysis, and problem-solving. Distinguishing between the three is nevertheless meaningful for crowdsourcing design in terms of the crowds to be targeted, evaluation, validation, and use of crowd outputs, and incentives models (Malone and Johns, 2011).

2.1.1. Ontology Engineering and Knowledge Base Curation

A number of projects in the Semantic Web area showcase the application of crowdsourcing techniques to collect new data to build ontologies and knowledge bases. Wikidata (Vrandečić and Krötzsch, 2014) is a community-oriented effort to create and curate a structured online knowledge base that is used by every language version of Wikipedia[1]. Other prominent examples include Freebase[2] and ontologies in the biomedical domain such as ICD (Tudorache et al., 2013a). There are a multitude of tasks which the crowd can contribute to in these cases, from defining classes, instances, class hierarchies, and relationships, to adding labels, documentation, and metadata to ontological primitives.

[1]https://www.wikipedia.org/

[2]https://www.freebase.com/

Celino et al. (2012a) developed a game to collect urban geo-spatial data. Noy et al. (2013) investigated the use of microtask platforms in ontology engineering comparing crowd workers with students and experts, while Hanika et al. (2014) proposed an extension of a popular ontology engineering environment with integrated GWAPs and microtask capabilities. Additionally, the crowd can become useful for understanding the different dimensions of culture, which is intrinsic to humans and needs to be reflected in semantic data. The crowd can also help in providing contextual knowledge, which is needed to adapt information to different target audiences (e. g. doctors and patients fora discussing rare diseases).

2.1.2. Validation and Enhancement of Knowledge

The diversity of data sources on the Web that contain different types of information and use different representations, naming conventions and natural languages introduces a challenging scenario for automatic methods to extract and process semantic data. Both in tasks related to ontology management and instance data (e. g. entity linking, RDB-to-RDF translation, fact checking and data interlinking), the quality and the amount of automatically generated knowledge can be improved by letting the crowd analyze, verify, correct or extend a particular aspect of an ontology or knowledge base.

Thaler et al. (2011) demonstrated how to gamify ontology alignment. Waitelonis et al. (2011) explored similar principles to curate DBpedia content. In microtask crowdsourcing, Demartini et al. (2012) worked on the identification of links between text entities and DBpedia URIs, while Sarasua et al. (2012) focused on the post-processing of ontology mappings generated by alignment algorithms. Kontokostas et al. (2013) proposed a contest to attract volunteers to assess Linked Data triples, which Acosta et al. (2013) combined with the use of microtasks. Simperl et al. (2013a) provided an overview of some of these and other examples in this area.

2.2. **Research Challenges**

Works such as (Kittur et al., 2013; Kern, 2014) discuss the main challenges in crowdsourcing research, including the design of crowdsourcing workflows, methods for task and roles assignment for contributors, quality control, and motivation and incentives. We find specific instantiations of these challenges in a Semantic Web context as well.

2.2.1. Task and Workflow Design

The selection and organization of tasks has a great impact on the success of any crowdsourcing endeavor. Decisions about task granularity, as well as the way information is presented to potential contributors can draw the line between a compelling and easy-to-understand task, and an obscure task (Moussawi and Koufaris, 2013). This is particularly true in microtask crowdsourcing, in which work is broken down in simple, routine tasks to be executed by the crowd in a matter of minutes and without particular training. Some Semantic Web tasks are naturally amenable to this type of divide-and-conquer strategy. For example, defining or validating mappings between two ontologies can be divided into smaller units of work, each referring to one pair of entities to be compared. In other scenarios (e. g. semantic annotation of large textual documents), tasks require a significant amount of contextual information to yield useful results, hence, it is not obvious how the annotation effort could be distributed to a large number of independent contributors (e.g., sentence level or paragraph level) while making sure that the overall meaning is not lost.

When it comes to the information to be displayed, a challenging aspect is to identify the minimum amount of context and domain knowledge that contributors need to have access to in order to accomplish the task correctly. For example, when crowdsourcing ontology engineering and lexicalization, the tasks should include some description of the classes, related and neighbouring entities, and relationships to be examined. In addition, one needs to create human-readable versions of the machine-readable data, either by using existing documentation or verbalization methods.

Finally, when tasks refer to an expert domain and not to popular knowledge, they have to be accompanied by specific examples and instructing methods. Traditionally, such examples have been created by requesters, but one could also imagine a scenario in which they are curated by the most engaged crowd contributors.

2.2.2. Using Multiple Crowdsourcing Genres

Aligning incentives is at the core of any socio-technical system that relies on human input and intervention (Kim, 2000; Kraut et al., 2012). Applications of design recommendations and guidelines from related disciplines such as gamification or online communities in Semantic Web contexts are largely unexplored (Simperl et al., 2013b). Studies in crowdsourcing have acknowledged that for example, volunteer-based crowdsourcing is able to collect humans with higher expertise than microtask crowdsourcing. However, research in our field has shown that crowdsourcing the verification of ontology relations via microtasks can also work in a highly specific domain like biomedicine (Mortensen et al., 2013b). Therefore, it is worth researching which crowdsourcing model (or which combination) is most suitable for each type of semantic management task, under which circumstances. A matching between a characterization of different types of tasks and the characterization of different crowds could help requesters in designing their systems.

The crowd has often been introduced as an alternative to the work carried out by experts (e. g. book translation (Minder and Bernstein, 2012)). However, there are Semantic Web tasks in which combining expert curation with crowd contributions might be the most effective approach, due to the knowledge-intensive aspect of the tasks. Whether the emphasis should be placed on the experts guiding crowds or the experts reviewing crowd contributions is still an open question.

2.2.3. Managing Hybrid Workflows

While with crowdsourcing we could, in theory, process manually large amounts of data, this comes at a cost in money and time. More efficient solutions are defined as hybrid crowd-powered approaches, which combine algorithmic and human-computation techniques (Bernstein, 2013). One of the topics to be researched in this direction is the interaction of the crowd with existing Semantic Web technologies and tools and the trade-offs of different variants of hybrid workflows in terms of accuracy, execution time, and costs. Going a step further, for crowdsourcing to establish itself in Semantic Web technology stacks and organizational processes, it will be essential to study paradigms and process models by which crowd computation could be added to existing workflows in a non-intrusive and systematic way.

2.2.4. Quality of Contributions

A number of methods have been proposed in crowdsourcing literature to assess and control the quality of crowd contributions (Quinn and Bederson, 2011; Kern et al., 2012; Schulze et al., 2012). Typically one distinguishes between manual (e. g. expert judgement, peer review) and automatic (e. g. majority voting) techniques and ensure that the overall approach remains cost-effective and yields useful results. Many of these techniques can be applied for Semantic Web tasks. However, it is yet to be determined which of them are more appropriate in each scenario, given the knowledge-intensive and often subjective nature of the questions to be solved. Novel frameworks for quality modeling and assessment are needed to reflect the richness of domain knowledge and insight that collective intelligence could bring into traditional computational processes, going beyond the rigid, Boolean view of ground truth used in many computer science areas. The work of Inel et al. (2014) around the CrowdTruth framework proposes a useful starting point. Methods that match ontologies to microtask design could also offer a helpful way to prune the space of possible solutions.

2.2.5. Finding and Managing a Suitable Crowd

Finding the right crowd is still an open challenge in crowdsourcing in general. Zogaj and Bretschneider (2014) argue that crowdsourcing projects are completed successfully and fast when certain tasks are distributed to experienced contributors who also enjoy handling the tasks. In our field, the first question that should be looked into in more detail is related to the set of cognitive skills that are required to accomplish specific types of Semantic Web tasks. Historically, our community has targeted domain experts and knowledge engineers to contribute to ontology and data management projects; in a crowdsourcing setting such assumptions cease to hold and new, more refined processes are needed to leverage the abilities of larger groups of people, who have only very limited or no insight at all in the technicalities of the Semantic Web. Once we have an understanding of this set of skills, we should devise new types of qualifications tests to identify the most suitable crowd for a given task and encourage participation. The work of Feldman and Bernstein (2014) proposes a set of cognitive tests that can be applied in the context of crowdsourcing. Equally important to qualification tests are e-Learning technologies which would be suitable for large-scale distributed scenarios to familiarize and instruct the crowd with the tasks to be performed and the underlying domain. Ipeirotis and Gabrilovich (2014) presented an approach for targeted volunteer crowdsourcing, which attracts expert users through ads that are strategically published in Web sites frequently visited by domain experts. This work, which might be applicable to other forms of crowdsourcing, shows that it is feasible to recruit the right people for a highly domain-specific task. Gathering and integrating available user information (after obtaining user consent) also becomes useful for finding the right crowd. Difallah et al. (2013) introduced a system in this direction, which proposes tasks to contributors depending on their user profile generated from information extracted from Facebook.

2.2.6. Understanding Social Dynamics

Social computing systems with massive human participation tend to aggregate a large set of users with different interests and behaviour. Understanding who the users are, the motivations behind their actions, they way they organize themselves and the way they accomplish tasks becomes crucial for the design of effective complex systems. This becomes particularly relevant in the context of collaborative crowdsourcing, in which contributions are created and edited by different agents in a cooperative manner. Tudorache et al. (2011, 2013b) adopted such approach in the context of ontology engineering, to enable the distributed definition of agreed-upon knowledge representations. The work of Falconer et al. (2011) on the analysis of activity logs suggests that users can play different roles (e. g. ontology expert, domain expert, content manager) in terms of editing patterns. Walk et al. (2014) observed the edit sequences in such processes. Strohmaier et al. (2013) investigated the way ontologies are collaboratively created analyzing dynamism, social aspects, lexis and behaviour. They found weak forms of collaboration among users and identified that the users with higher degree of contribution were the most central users. These findings shed light on the way users act. However, there is still much to research about emerging social coordination mechanisms. For example, an open question is whether recommendation algorithms could help in the collaborative production of ontologies.

3. SEMANTIC WEB FOR CROWDSOURCING

We identify three core contributions that the Semantic Web can offer to crowdsourcing tools: first, machine-processable semantics facilitates the formal, explicit specification of the crowdsourcing domain, with all its components. Second, Linked Data standards and protocols facilitate information integration and reuse across crowdsourcing platforms and experiments. Third, reasoning could enhance the capabilities of specific crowdsourcing-related methods.

3.1. Scenarios

We describe the aforementioned three scenarios.

3.1.1. Knowledge Representation

The great majority of crowdsourcing platforms offer only limited means to match tasks to contributors. Basic crowd filtering features offered in such platforms are geographic location or high-level reputation scores, but essential aspects such as contributors' skills and knowledge are often not available. Even when such information is in place (e. g. in macrotask environments), it is primarily as unstructured text descriptions. For the same reasons, contributors are not optimally assisted in finding those tasks that best match their preferences. A common and structured representation of the crowdsourcing domain could be used to implement more advanced search features. The use of semantic technologies to realize this representation means improved matchmaking capabilities between the information needs of the users and the pool of available tasks and human resources. Platforms for software testing like testcloud and testbirds provide an example of how contributors may be selected based on their accomplished work and self-specified preferences (Zogaj et al., 2014).

Such semantic descriptions, together with other Linked Open Data sets, could be exploited to adjust crowdsourcing workflows. For example, information about labor regulations and statistical governmental resources could be taken into account to ensure fairness for both contributors and requesters in a paid crowdsourcing environment. This automatic analysis could complement more in-depth manual certifications granted by official authorities. Semantic Web technologies could also be used to publish the data that has been generated from crowdsourcing experiments, as a first step towards reproducibility of research results.

Moreover, a widely agreed conceptualization of the crowdsourcing landscape would have positive effects on the further development of the field and all its stakeholders, making communication more effective and enabling comparative studies of the research field.

3.1.2. Data Integration

The use of Linked Data as an integration technology means that new, external sources of information could be pulled in to inform decisions that requesters need to be made in the crowdsourcing process. For example, ontology terms can be used to automatically build qualification questions to assess the knowledge of contributors in particular domains (e. g. the NCI thesaurus can be used to test the knowledge on cancer treatments). This can considerably ease the construction of such questions, which currently need to be defined manually by requesters. Another example in this direction is the use of data on the Linked Open Data cloud to provide additional context about the data in the crowdsourcing tasks.

Data generated in one platform could be reused and integrated with other sources to increase its value. The same is true about crowdsourcing processes. Using Semantic Web standards and principles would make this knowledge accessible to machines and facilitate a more systematic and flexible design of crowdsourcing projects, which will be able to build and reuse arbitrarily complex combinations of relevant components and services produced by a variety of parties. The Semantic Web community has been very appreciative of this type of model-driven approach to software engineering, proposing frameworks such as Semantic Web services (Fensel et al., 2011), Linked APIs (Krummenacher et al., 2010), or semantic workflows (Cardoso and Sheth, 2003) to describe computational processes in a way that facilitates their automatic discovery, matchmaking and composition. These formalisms and technology should be revisited to assess their use for the representation of the functional and non-functional characteristics of various crowdsourcing platforms and tools.

3.1.3. Automatic Reasoning

The Semantic Web community has developed a rich portfolio of algorithms and tools to infer implicit knowledge and identify inconsistencies in semantically annotated data. One of the uses of this powerful feature is the generation of feedback to crowd contributions, possibly in combination with manual editing by the requester. This would make task management more effective not just because parts of the feedback cycle would be automated, but most importantly, because it would help contributors to get better at their tasks and motivate them to continue (Kraut et al., 2012).

An automatic analysis of the consistency of crowd responses, for instance, by taking into account specific properties and constraints that hold in the task domain, could assist existing quality control methods (Sheng et al., 2008). Similarly, based on the responses obtained from the crowd, one could imagine a system in which complex requester workflows are adaptively created following insights from a background knowledge base. Once validated, these responses could enrich existing data sources and derive new knowledge facts.

3.2. **Research Challenges**

In order to realize the scenarios just discussed and develop semantically enabled crowdsourcing technology more research is needed to provide answers to the following open questions.

3.2.1. Defining Vocabularies or Ontologies

The Semantic Web community has published hundreds of ontologies covering different domains[3]. While it is possible (and encouraged) to reuse existing and widely deployed ontologies and vocabularies, the crowdsourcing context has its own information needs. Therefore, new ontologies should be engineered to enable the representation of knowledge on contributors, requesters, tasks, workflows, quality control and resulting data. As a community, we need to reach a common understanding of the commonalities and differences between different forms of crowdsourcing, design vocabularies capturing these findings, and share them widely. The vocabularies developed for provenance management and the social Semantic Web become of particular interest for this purpose. Recent proposals include the work of Celino (2013) on the Human Computation Ontology, which enables the annotation of objects resulting from human computation processes and is mapped to PROV[4]. Sarasua and Thimm (2014) propose Crowd Work CV, an ontology to capture crowd workers' and requesters' information across different crowdsourcing platforms. The CrowdTruth framework proposes a useful schema for annotating the provenance of crowdsourced data (Inel et al., 2014). A second challenge will be to identify the most appropriate level of granularity to record activities and their results to meet the requirements of various use cases (data aggregation, quality control, tasks assignment, personalization etc.)

3.2.2. Connecting Linked Data and Crowdsourced Data

As discussed earlier, using Semantic Web standards and principles for data publication and use facilitates seamless integration among data sources and processes to create more advanced crowdsourcing pipelines. For this powerful idea to materialize, one needs tools that identify links between the individual data sets (e. g. annotating results from the crowds with external Linked Data URIs) or discover relevant crowd computing functionality. Connecting crowdsourced Linked Data to the LOD cloud[5] poses specific challenges: for instance, one could use this form of disambiguation of crowd results to identify incorrect answers or to resolve disagreement cases which are just a signal for poor accuracy rather than diversity of viewpoints (Inel et al., 2014). Offering a full-fledged semantic workflow solution remains one of the greatest challenges of Semantic Web research; narrowing the domain down to crowdsourcing scenarios might reduce its complexity. However, it would still require tools to create useful semantic descriptions and automatically record or mine provenance information. In addition, relevant stakeholders need to acknowledge the benefits and commit to publishing data and software for the greater community to use, which is a problem that can not be solved just at the technology level.

---

[3] LOV portal http://lov.okfn.org/dataset/lov/
[4] PROV http://www.w3.org/TR/prov-o/
[5] LOD http://lod-cloud.net/

3.2.3. Generating Specific Knowledge for Automatic Reasoning

A primary use of Semantic Web reasoning is to identify inconsistencies in knowledge bases; this could be contradicting information obtained from the crowd, or new facts that are not consistent with background knowledge. This will require non-traditional methods to gain insight from crowd data, which is noisy, uncertain, and aggregated from multiple contributors.

Another application is policy management for ethical and fair crowdsourcing. This is challenging primarily because of the complexity of the domain to be modeled, which requires sophisticated reasoning algorithms that are able to deal with multiple contexts.

4. GUIDELINES

From the discussions around the scenarios introduced in the previous sections we compiled a list of nine guidelines which should be taken into account when conducting research in this area:

Crowd contributors are not Semantic Web experts To be truly effective, attempts to crowdsource Semantic Web tasks should assume that the people who will take on these tasks have no technical background. This requires, as crucial elements, human-readable descriptions of Semantic Web resources (no RDF, URIs, HTML tags etc), contextual information (e.g., adjacent nodes in the knowledge graph), and many representative and well-made educational examples. The use of technical terms such as *triple*, *ontology*, *class* or *Linked Data* should be avoided as well, as this nomenclature is not only irrelevant for the completion of the task, but may confuse non-experts.

Use machine-processable semantics to describe crowd processes Semantically representing information can enable the inference of new and implicit facts. The semantic annotation of the different aspects of crowd processes (e.g., contributors, requesters, tasks and their outcomes) can lead to a better automatic management of such processes. In addition, formal descriptions of legal and ethical aspects could help identifying inconsistencies and prevent the violation of pre-defined constraints (e.g., compliance with labor laws).

Use machine-processable semantics to improve task design and assignment The fact that we are using Semantic Web data allows us to undertake specific optimizations in the choice of tasks to be assigned to the crowd. This includes, for example, the identification of most suitable answer choices, the detection of inconsistencies in crowd answers, or the ordering of tasks in bundles to improve crowd performance.

Define a framework to capture the meaning of uncertain and subjective knowledge Many tasks in semantic content creation use open online sources that contain uncertain, inconsistent, or subjective knowledge. These characteristics need to be reflected in the design of the tasks and in the way results are assessed and rewarded.

Acknowledge agreement and disagreement Traditionally, crowdsourced data aggregation techniques have focused on consensus, discarding crowd contributions that differ from those provided by the majority of the crowd. However, disagreement may be an indicator of low-quality data. Moreover, in many scenarios acknowledging different points of view may be enriching and provide a definition of the context.

Open crowdsourced data together with machine-readable metadata Research data created or curated through crowdsourcing should be made openly available to avoid 'wasting' valuable human resources by repeating unnecessary experiments. From an operational point of view, this includes formats and technologies that facilitate reuse, in particular metadata describing both the data, and the processes by which it was obtained. The Semantic Web community has developed a number of vocabularies which could provide a useful starting point: DCAT[6], VoiD[7] and DDI[8] for data set description, PROV-O[9] for provenance information and the RDF version of Creative Commons[10] licenses for reporting licensing information.

Be open about your crowdsourcing project Share as much information as possible about the goal of the crowdsourcing process with workers. Inform them about the reason for the crowdsourcing task they have to solve, the provenance of the data they need to process and the way their contribution (e.g. crowdsourced data) will be (re)used.

Share your work and findings with the community Because the research field at the intersection of crowdsourcing and Semantic Web technologies is in its early stages, every definition of a new use case and every empirical evaluation will provide the research community with a better understanding of the domain and the methods to be used. For this reason, researchers should share not only the collected data, but also the implemented **algorithms**, any **other resources** they develop for their works and the **lessons they learned**. Specially, since our community values technologies for open access, interoperability and data reuse. We should take advantage of available open source platforms (e. g. GitHub, Sourceforge etc.), as well as data hosting and cataloging infrastructures (e. g. CKAN, the LOD cloud).

5. CONCLUSIONS

Due to the novelty and continuous evolution of these technologies, there is still much to explore, study and understand in the interdisciplinary field that we cover in this manifesto. With this document we aim at shedding some light on the opportunities that further research could lead to.

Recent success stories in ontology engineering, Linked Data management and semantic annotation of Web content show that there is a promising future for the use of crowdsourcing techniques in Semantic Web tasks. The use of Semantic Web methods in crowdsourcing environments has been investigated to a smaller extent, but as we described here, it can be equally beneficial.

We encourage the research community to design new methodologies and best practices, build infrastructure and design algorithms to combine human and machine computation for and with the Web of Data. All this, applying our own principles of shareable, reusable and open knowledge and remembering that with crowdsourcing we are opening the Semantic Web to a wider audience.

---

[6]DCAT vocabulary http://www.w3.org/TR/vocab-dcat/
[7]VoiD vocabulary http://www.w3.org/TR/void/
[8]DDI vocabulary http://rdf-vocabulary.ddialliance.org/discovery.html
[9]PROV-O Ontology http://www.w3.org/TR/prov-o/
[10]CC in RDF http://creativecommons.org/ns

6. ACKNOWLEDGMENTS

We thank the participants of the *Dagstuhl Seminar 14282* for their contributions and interesting discussions: Maribel Acosta, Sofia Angeletou, Lora Aroyo, Irene Celino, Philippe Cudré-Mauroux, Roberta Cuel, Gianluca Demartini, Michael Feldman, Yolanda Gil, Carole Goble, Robert Kern, Atsuyuki Morishima, Valentina Presutti, Marta Sabou, Harald Sack, Markus Strohmaier, Gerd Stumme, Tania Tudorache, Maja Vukovic, Christopher A. Welty and Marco Zamarian.

7. REFERENCES

Acosta, M, Simperl, E, Flöck, F, Vidal, M.-E, and Studer, R. (2015). RDF-Hunter: Automatically Crowdsourcing the Execution of Queries Against RDF Data Sets. *Arxiv preprint arXiv:1503.02911* (2015).

Acosta, M, Zaveri, A, Simperl, E, Kontokostas, D, Auer, S, and Lehmann, J. (2013). Crowdsourcing Linked Data Quality Assessment. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II*. 260–276.

Atzmueller, M, Becker, M, Kibanov, M, Scholz, C, Doerfel, S, Hotho, A, Macek, B.-E, Mitzlaff, F, Mueller, J, and Stumme, G. (2014). Ubicon and its Applications for Ubiquitous Social Computing. *New Review of Hypermedia and Multimedia* 1, 20 (2014), 53–77. DOI:http://dx.doi.org/10.1080/13614568.2013.873488

Baker, T, Noy, N, Swick, R, and Herman, I. (2012). Semantic Web Case Studies and Use Cases. (2012). http://www.w3.org/2001/sw/sweo/public/UseCases/

Bernstein, A. (2012). The global brain semantic web Interleaving Human-Machine knowledge and computation. In *Workshop on What will the Semantic Web Look Like 10 Years From Now? at ISCW 2012, Boston, MA.*

Bernstein, A, Leimeister, J. M, Noy, N, Sarasua, C, and Simperl, E. (2014). Crowdsourcing and the Semantic Web (Dagstuhl Seminar 14282). *Dagstuhl Reports* 4, 7 (2014), 25–51. DOI:http://dx.doi.org/10.4230/DagRep.4.7.25

Bernstein, M. S. (2013). Crowd-Powered Systems. *KI* 27, 1 (2013), 69–73.

Bigham, J. P, Bernstein, M. S, and Adar, E. (2015). *HCI Human-Computer Interaction and Collective Intelligence*.

Blohm, I, Leimeister, J. M, and Krcmar, H. (2013). Crowdsourcing: How to Benefit from (Too) Many Great Ideas. *MIS Quarterly Executive* 12, 4 (2013).

Boudreau, K. J and Lakhani, K. R. (2013). Using the crowd as an innovation partner. *Harvard business review* 91, 4 (2013), 60–69.

Burke, J. A, Estrin, D, Hansen, M, Parker, A, Ramanathan, N, Reddy, S, and Srivastava, M. B. (2006). Participatory sensing. *Center for Embedded Network Sensing* (2006).

Cardoso, J and Sheth, A. (2003). Semantic e-workflow composition. *Journal of Intelligent Information Systems* 21, 3 (2003), 191–225.

Celino, I. (2013). Human Computation VGI Provenance: Semantic Web-Based Representation and Publishing. *IEEE T. Geoscience and Remote Sensing* 51, 11 (2013), 5137–5144.

Celino, I, Cerizza, D, Contessa, S, Corubolo, M, Dell'Aglio, D, Valle, E. D, and Fumeo, S. (2012)a. Urbanopoly - A Social and Location-Based Game with a Purpose to Crowdsource Your Urban Data. In *2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Confernece on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3-5, 2012*. 910–913. DOI:http://dx.doi.org/10.1109/SocialCom-PASSAT.2012.138

Celino, I, Contessa, S, Corubolo, M, Dell'Aglio, D, Valle, E. D, Fumeo, S, and Krüger, T. (2012)b. Linking Smart Cities Datasets with Human Computation - The Case of UrbanMatch. In *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part II*. 34–49. DOI:http://dx.doi.org/10.1007/978-3-642-35173-0_3

Demartini, G, Difallah, D. E, and Cudré-Mauroux, P. (2012). ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*. 469–478. DOI:http://dx.doi.org/10.1145/2187836.2187900

Demartini, G, Trushkowsky, B, Kraska, T, and Franklin, M. J. (2013). CrowdQ: Crowdsourced Query Understanding. In *CIDR 2013, Sixth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 6-9, 2013, Online Proceedings*.

Difallah, D. E, Demartini, G, and Cudré-Mauroux, P. (2013). Pick-a-crowd: tell me what you like, and i'll tell you what to do.. In *WWW*, Daniel Schwabe, Virgílio A. F. Almeida, Hartmut Glaser, Ricardo A. Baeza-Yates, and Sue B. Moon (Eds.). International World Wide Web Conferences Steering Committee / ACM, 367–374.

Doan, A, Ramakrishnan, R, and Halevy, A. Y. (2011). Crowdsourcing systems on the World-Wide Web. *Commun. ACM* 54, 4 (2011), 86–96.

Falconer, S, Tudorache, T, and Noy, N. F. (2011). An analysis of collaborative patterns in large-scale ontology development projects. In *Proceedings of the sixth international conference on Knowledge capture*. ACM, 25–32.

Feldman, M and Bernstein, A. (2014). Cognition-based Task Routing:Towards Highly-Effective Task-Assignments in Crowdsourcing Settings. In *35th International Conference on Information Systems (ICIS 2014)*. s.n., Auckland, New Zealand.

Fensel, D, Facca, F. M, Simperl, E, and Toma, I. (2011). *Semantic web services*. Springer.

Ferrucci, D. A, Brown, E. W, Chu-Carroll, J, Fan, J, Gondek, D, Kalyanpur, A, Lally, A, Murdock, J. W, Nyberg, E, Prager, J. M, Schlaefer, N, and Welty, C. A. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31, 3 (2010), 59–79.

Hanika, F, Wohlgenannt, G, and Sabou, M. (2014). The uComp Protégé Plugin: Crowdsourcing Enabled Ontology Engineering. *Semantic Web Journal (EKAW2014)* (2014).

Hitzler, P, Krötzsch, M, and Rudolph, S. (2009). *Foundations of Semantic Web Technologies*. CRC Press, Boca Raton, FL.

Howe, J. (2008). *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Publishing Group.

Inel, O, Khamkham, K, Cristea, T, Dumitrache, A, Rutjes, A, van der Ploeg, J, Romaszko, L, Aroyo, L, and Sips, R.-J. (2014). CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data. In *The Semantic Web–ISWC 2014*. Springer, 486–504.

Ipeirotis, P. G and Gabrilovich, E. (2014). Quizz: targeted crowdsourcing with a billion (potential) users. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*. 143–154. DOI:http://dx.doi.org/10.1145/2566486.2567988

Kern, R. (2014). Dynamic Quality Management for Cloud Labor Services. Methods and Applications for Gaining Reliable Work Results with an On-Demand Workforce. Series: Lecture Notes in Business Information Processing, Vol. 192. Springer.

Kern, R, Thies, H, Zirpins, C, and Satzger, G. (2012). Dynamic and Goal-Based Quality Management for Human-Based Electronic Services. *Int. J. Cooperative Inf. Syst.* 21, 1 (2012), 3–29.

Kim, A. J. (2000). *Community building on the Web: Secret strategies for successful online communities*. Addison-Wesley Longman Publishing Co., Inc.

Kittur, A, Nickerson, J. V, Bernstein, M. S, Gerber, E, Shaw, A. D, Zimmerman, J, Lease, M, and Horton, J. (2013). The future of crowd work.. In *CSCW*, Amy Bruckman, Scott Counts, Cliff Lampe, and Loren G. Terveen (Eds.). ACM, 1301–1318.

Kontokostas, D, Zaveri, A, Auer, S, and Lehmann, J. (2013). TripleCheckMate: A Tool for Crowdsourcing the Quality Assessment of Linked Data. In *Knowledge Engineering and the Semantic Web - 4th International Conference, KESW 2013, St. Petersburg, Russia, October 7-9, 2013. Proceedings*. 265–272.

Kraut, R. E, Resnick, P, Kiesler, S, Burke, M, Chen, Y, Kittur, N, Konstan, J, Ren, Y, and Riedl, J. (2012). *Building successful online communities: Evidence-based social design*. MIT Press.

Krummenacher, R, Norton, B, and Marte, A. (2010). Towards linked open services and processes. In *Future Internet-FIS 2010*. Springer, 68–77.

Leimeister, J. M, Huber, M, Bretschneider, U, and Krcmar, H. (2009). Leveraging Crowdsourcing: Activation-Supporting components for IT-based ideas competition. *Journal of Management Information Systems (JMIS)* 26, 1 (2009), 197–224. http://pubs.wi-kassel.de/wp-content/uploads/2013/03/JML_145.pdf 138 (22-09).

Malone T. Laubacher, R and Johns, T. (2011). The Big Idea: The Age of Hyperspecialization. *Harvard business review July 2011* (2011).

Minder, P and Bernstein, A. (2012). How to translate a book within an hour: towards general purpose programmable human computers with crowdlang. In *Proceedings of the 3rd Annual ACM Web Science Conference*. ACM, 209–212.

Morishima, A, Shinagawa, N, Mitsuishi, T, Aoki, H, and Fukusumi, S. (2012). CyLog/Crowd4U: A Declarative Platform for Complex Data-centric Crowdsourcing. *PVLDB* 5, 12 (2012), 1918–1921. http://vldb.org/pvldb/vol5/p1918_atsuyukimorishima_vldb2012.pdf

Mortensen, J, Alexander, P. R, Musen, M. A, and Noy, N. F. (2013)a. Crowdsourcing Ontology Verification. In *Proceedings of the 4th International Conference on Biomedical Ontology, ICBO 2013, Montreal, Canada, July 7-12, 2013*. 40–45. http://ceur-ws.org/Vol-1060/icbo2013_submission_51.pdf

Mortensen, J, Musen, M. A, and Noy, N. F. (2013)b. Crowdsourcing the Verification of Relationships in Biomedical Ontologies. In *AMIA 2013, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 16-20, 2013*. http://knowledge.amia.org/amia-55142-a2013e-1.580047/t-09-1.582024/f-009-1.582025/a-345-1.582084/a-363-1.582079

Moussawi, S and Koufaris, M. (2013). The Crowd on the Assembly Line: Designing Tasks for a Better Crowdsourcing Experience.

(2013).

Noy, N. F, Mortensen, J, Musen, M. A, and Alexander, P. R. (2013). Mechanical turk as an ontology engineer?: using microtasks as a component of an ontology-engineering workflow.. In *WebSci*, Hugh C. Davis, Harry Halpin, Alex Pentland, Mark Bernstein, and Lada A. Adamic (Eds.). ACM, 262–271.

Quinn, A. J and Bederson, B. B. (2011). Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Raddick, M. J, Bracey, G, Carney, K, Gyuk, G, Borne, K, Wallin, J, Jacoby, S, and Planetarium, A. (2009). Citizen science: status and research directions for the coming decade. *AGB Stars and Related Phenomenastro 2010: The Astronomy and Astrophysics Decadal Survey* (2009), 46P.

Raimond, Y, Ferne, T, Smethurst, M, and Adams, G. (2014). The BBC World Service Archive Prototype. *Web Semantics: Science, Services and Agents on the World Wide Web* 27, 1 (2014). http://www.websemanticsjournal.org/index.php/ps/article/view/378

Sarasua, C, Simperl, E, and Noy, N. F. (2012). CrowdMap: Crowdsourcing Ontology Alignment with Microtasks. In *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I*. 525–541. DOI:http://dx.doi.org/10.1007/978-3-642-35176-1_33

Sarasua, C and Thimm, M. (2014). Crowd Work CV: Recognition for Micro Work. In *Proceedings of the 3rd International Workshop on Social Media for Crowdsourcing and Human Computation (SoHuman'14)*.

Schulze, T, Krug, S, and Schader, M. (2012). Workers' Task Choice in Crowdsourcing and Human Computation Markets.. In *ICIS*. Association for Information Systems.

Shadbolt, N, Hall, W, and Berners-Lee, T. (2006). The semantic Web revisited. *Intelligent Systems, IEEE* 21, 3 (2006), 96–101.

Shadbolt, N. R, Smith, D. A, Simperl, E, Kleek, M. V, Yang, Y, and Hall, W. (2013). Towards a classification framework for social machines. In *22nd International World Wide Web Conference, WWW'13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*. 905–912.

Sheng, V, Provost, F, and Ipeirotis, P. G. (2008). Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, 614–622. http://archive.nyu.edu/bitstream/2451/25882/4/kdd2008.pdf isbn = 978-1-60558-193-4, location = Las Vegas, Nevada, USA, doi = http://doi.acm.org/10.1145/1401890.1401965,.

Simperl, E, Acosta, M, and FlÃűck, F. (2013)a. Knowledge Engineering via Human Computation. In *Handbook of Human Computation*, Pietro Michelucci (Ed.). Springer New York, 131–151. DOI:http://dx.doi.org/10.1007/978-1-4614-8806-4_13

Simperl, E, Cuel, R, and Stein, M. (2013)b. Morgan & Claypool Publishers.

Singhal, A. (2012). Introducing the Knowledge Graph: things, not strings. (2012).

Siorpaes, K and Simperl, E. (2010). Human intelligence in the process of semantic content creation. *World Wide Web* 13, 1-2 (2010), 33–59.

Strohmaier, M, Walk, S, Pöschko, J, Lamprecht, D, Tudorache, T, Nyulas, C, Musen, M. A, and Noy, N. F. (2013). How ontologies are made: Studying the hidden social dynamics behind collaborative ontology engineering projects. *Web Semantics: Science, Services and Agents on the World Wide Web* 20 (2013), 18–34.

Thaler, S, Simperl, E. P. B, and Siorpaes, K. (2011). SpotTheLink: playful alignment of ontologies. In *Proceedings of the 2011 ACM Symposium on Applied Computing (SAC), TaiChung, Taiwan, March 21 - 24, 2011*. 1711–1712. DOI:http://dx.doi.org/10.1145/1982185.1982542

Tudorache, T, Nyulas, C, Noy, N. F, and Musen, M. A. (2013)a. Using Semantic Web in ICD-11: Three Years Down the Road. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II*. 195–211.

Tudorache, T, Nyulas, C, Noy, N. F, and Musen, M. A. (2013)b. WebProtégé: A collaborative ontology editor and knowledge acquisition tool for the web. *Semantic web* 4, 1 (2013), 89–99.

Tudorache, T, Nyulas, C, Noy, N. F, Redmond, T, and Musen, M. A. (2011). iCAT: A Collaborative Authoring Tool for ICD-11. In *Workshop âĂIJOntologies come of Age in the Semantic WebâĂİ(OCAS2011) 10 th International Semantic Web Conference Bonn, Germany, October 24, 2011*. 72.

von Ahn, L and Dabbish, L. (2008). Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.

Vrandečić, D and Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

Vukovic, M and Bartolini, C. (2010). Towards a Research Agenda for Enterprise Crowdsourcing.. In *ISoLA (1) (Lecture Notes in Computer Science)*, Tiziana Margaria and Bernhard Steffen (Eds.), Vol. 6415. Springer, 425–434.

Waitelonis, J, Ludwig, N, Knuth, M, and Sack, H. (2011). WhoKnows? Evaluating linked data heuristics with a quiz that cleans up DBpedia. *Interact. Techn. Smart Edu.* 8, 4 (2011), 236–248.

Walk, S, Singer, P, Strohmaier, M, Tudorache, T, Musen, M. A, and Noy, N. F. (2014). Discovering beaten paths in collaborative ontology-engineering projects using markov chains. *Journal of biomedical informatics* 51 (2014), 254–271.

Zogaj, S and Bretschneider, U. (2014). Analyzing Governance Mechanisms for Crowdsourcing Information Systems: A Multiple Case Analysis. (2014).

Zogaj, S, Bretschneider, U, and Leimeister, J. M. (2014). Managing crowdsourced software testing: a case study based insight on the challenges of a crowdsourcing intermediary. *Journal of Business Economics* 84, 3 (2014), 375–405.