Incentives, Motivation, Participation, Games: Human Computation for Linked Data

Katharina Siorpaes[1] and Elena Simperl[1,2]

[1] STI, University of Innsbruck, Austria, [2] AIFB, Karlsruhe Institute of Technology, Germany
katharina.siorpaes@sti2.at, elena.simperl@kit.edu

Abstract. Various tasks in publishing and maintaining Linked Data require human contribution at several ends. In this paper, we discuss the role of human intelligence in data interlinking. This justifies the necessity of incentive models, motivation mechanisms, and applying the paradigm of human computation to the interlinking process. To conclude, we give examples for using human computation for interlinking by summarizing "games with a purpose" that address tasks in data interlinking.

1. Motivation

The Web of Data, also described as a *web of things in the world, described by data on the Web* " [1], is the result of the process of publishing linked data on the Web. Such Linked Data, a global data space, enables more comprehensive answers to queries over aggregated data. Machine-readable data with explicitly defined meaning can be consumed by machines in order to provide improved access to various information sources leading to an enhanced user experience.

Along with four principles for linked data, the basic process of publishing data has been defined in three main steps [1]. Even though substantial tool support for these steps has already been developed, the process still requires human contribution at several ends. This also holds for several other tasks in the area of knowledge acquisition, such as finding free tags describing images, certain problems in ontology alignment, or modeling domains [2].

As human intervention is required in many areas of knowledge acquisition, this labor must be motivated by extrinsic or intrinsic incentives. Several Web 2.0 applications, such as del.icio.us, Amazon Mechanical Turk, or Wikipedia, demonstrate successful implementation of various motivation mechanisms.

The idea of Human Computation [4] is that tasks that are trivial for humans but not solvable for computer programs are solved by channeling human labor. One example for this are CAPTCHAs [4], another line of work are "games with a purpose" [3].

Still: even if a game addressing a knowledge acquisition task is available, it is by no means a guarantee for community involvement and the generation of large amounts of data. Designing "games with a purpose" is a tricky task, as gaming fun, data acquisition, and data quality concerns must be considered and well balanced.

We argue that it is necessary to address incentive models and motivation mechanisms to involve human users in the interlinking process, not only when publishing but also with respect to maintenance of Linked Data.

In this paper, we first delineate why human contribution is required at several ends in the Linked Data lifecycle. We then describe games - as a form of incentive and motivation mechanism - that already address interlinking tasks.

2. Human Intelligence in Data Interlinking

The main tasks that have to be performed in order to publish data as Linked Data are (i) to assign consistent URIs to data published, (ii) to generate links, and (iii) to publish metadata allowing further exploration and discovery of relevant datasets. Methods and tools were developed for automating steps 1 and 3, such as D2R or Virtuoso. We therefore argue that the issue of link generation is the major, most challenging problem that requires human attention.

The main problem, which arises, is the issue of finding the matching concepts in datasets to be interlinked and to name the relationships between the interlinked concepts (using defined relations such as owl:sameAs, rdfs:seeAlso, foaf:birthPlace, foaf:homeTown or others).

Several approaches exist for semantically linking data: RDF links can be set manually - supported by a set of tools including URI search and recommendation engines such as Uriqr, Sindice, or MOAT[1].

While currently available interlinking algorithms yield good results for textual resources the question arises if the quality of the links can be increased. In our understanding this is mainly possible through the utilization of human power.

The paradigm of Human Computation and lessons learnt from Web 2.0, where users collaboratively create content and metadata should be applied to Linked Data. In this case, annotations are based on semantic links, that is, RDF properties.

“User Contributed Interlinking” [5] denotes this principle way of interlinking resources on the Web of Data. The crucial steps in interlinking are identifying a target dataset, a target link, and choosing a link predicate. While the first step is of technical nature and can be easily automated and the second step essentially is annotation of resources, the last three steps require sophisticated methods or human intelligence. The linking phase focuses on the final three steps of the process and is the main focus of methods for interlinking.

1 Identify target dataset

The choice of the target dataset involves knowledge of available datasets, such as DBPedia, Geonames, Freebase, and their domain and focus. In many cases, these datasets might also be overlapping. However, for all datasets, reliable and

[1] http://virtuoso.openlinksw.com, http://www.w3.org/TR/skos-reference/, http://dev.uriqr.com/, http://www.sindice.com/, http://moat- project.org

comprehensive descriptions of their scope and purpose are available. This allows for a higher degree of automation.

2 Identify link target

Many datasets can be large and thus identifying the relevant link target can require machine support. However, finding the right link target might be tricky.

3 Select link predicate

The final step is the most challenging one as it describes the type of relationship that exists between two nodes.

When discussing manual labor in interlinking, just like in annotation, the type of content that is interlinked must be considered. For interlinking methods, we are aware of one survey published by Scharffe and Euzenat (2009)[2] that also investigated the degree of automation interlinking tools can support. The authors analyzed six tools and conducted interviews with the developers. The result was that except one system all tools where classified as semi-automatic, all requiring human intervention. The automatic method only worked for a specific domain. For the area of multimedia interlinking [8], the degree of automation reduces. However, this is tightly coupled to the quality and availability of annotations of the multimedia content.

3. Human Computation for Linked Data

In the previous section, we briefly summarized why human involvement is required for publishing, maintaining, and consuming Linked Data. One form of human computation and incentives are games that hide the abstract task behind an entertaining user interface (and story). In this section, we shortly summarize three games that aim at using or producing linked data.

The GuessWhat! game [7] is a multi-player online game that leverages the “games with a purpose” paradigm and Linked Open Data as a data source in order to build formal vocabularies (or domain ontologies). In the game, players are confronted with class expressions such as fruit AND yellow AND grows on tree automatically generated from Linked Open Data. The players have to invent a suitable class name (banana or lemon, for example) as fast as possible. The player with the highest number of plausible class labels wins the game.

SpotTheLink[3] [6], the latest release of the OntoGame framework, is a game that allows for the definition of mappings between ontologies as part of a collaborative game experience. In the game, players have to agree on the type of relationship between two concepts (or entities). The background is that a multitude of approaches to match, merge and integrate ontologies and to interlink RDF data sets have been

[2] http://melinda.inrialpes.fr/systems.html

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

proposed. While advances in this area cannot be contested, it is equally true that full automation of the ontology-alignment process is far from being feasible, and human intervention is often indispensable - mainly for bootstrapping the underlying methods and for validating and enhancing their results.

TubeLink[4] (Figures 1 and 2) is another game of the OntoGame series that will be published in early 2011. The idea is to use data on the Web to bootstrap the process of video annotation. In the game, players have to choose suitable tags describing contents of videos. These tags are really pieces of information that help methods in the background to choose appropriate data from the Linked Open Data cloud. At some point in time, players' input is used to choose an appropriate dataset, another time it might be used for selecting instances of that set. However, all this complexity is well hidden from the player.

Figure 1 TubeLink (upcoming: www.insemtives.eu or www.ontogame.org)

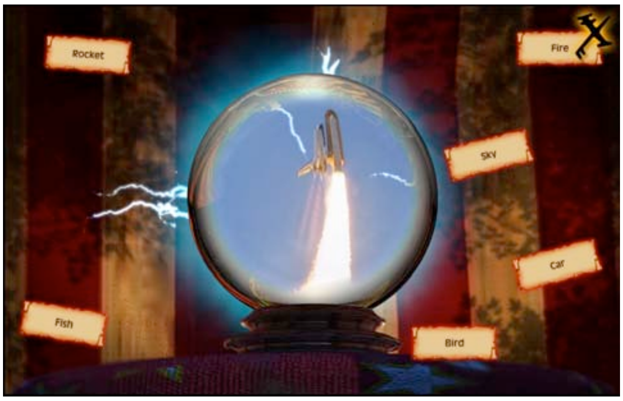

Figure 2 TubeLink video annotation (upcoming: www.insemtives.eu or www.ontogame.org)

[4] TubeLink is a part of the INSEMTIVES OntoGame series and will be published soon. Check back at www.insemtives.eu or www.ontogame.org

4. Conclusion

Publishing, maintaining, consuming linked data and thus contributing to a Web of Data involves several tasks that are partly dependent on human intelligence and intervention. We discussed that many methods for interlinking are semi-automatic and thus require user intervention. Therefore, it is necessary to address incentive models and motivation mechanisms to involve human users in the interlinking process, including publishing, maintenance, and consumption. One example of incentives and applying the paradigm of human computation are "games with a purpose". We described three example games that somehow address Linked (Open) Data.

Acknowledgments. This work has been funded by the EU FP7 project INSEMTIVES – Incentives for Semantics (www.insemtives.eu).

References

1. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data – The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), 2009.
2. Katharina Siorpaes and Elena Simperl: Human Intelligence in the Process of Semantic Content Creation, World Wide Web Journal (WWW), Volume 13, Issue 1-2, March 2010.
3. Luis von Ahn and Laura Dabbish: Designing games with a purpose. Communications of the ACM 51(8), 58–67, 2008.
4. Luis von Ahn: Human Computation, K-CAP '07 Proceedings of the 4th international conference on Knowledge capture, ACM, 2007.
5. Michael Hausenblas, Raphael Troncy, Tobias Bürger, and Yves Raimond "Interlinking Multimedia: How to Apply Linked Data Principles to Multimedia Fragments" In: Proceedings of Linked Data on the Web (LDOW2009), co-located with the 18th International World Wide Web Conference (WWW2009), Madrid, Spain, 2009.
6. Stefan Thaler, Elena Simperl, Katharina Siorpaes: SpotTheLink: Playful Alignment of Ontologies, 26th Symposium On Applied Computing (SAC'11), TaiChung, Taiwan, March 21-25, 2011. (upcoming)
7. Thomas Markotschi and Johanna Völker. GuessWhat?! - Human Intelligence for Mining Linked Data. Proceedings of the Workshop on Knowledge Injection into and Extraction from Linked Data (KIELD) at the International Conference on Knowledge Engineering and Knowledge Management (EKAW), 2010.