Democratizing Open Data

Alvaro Graves
Tetherless World Constellation
Department of Cognitive Science
Rensselaer Polytechnic Institute
gravea3@rpi.edu

Abstract

Progress has been made in terms of the availability of Open Data initiatives in dozens of universities, governments and organizations. For example, Open Government Data (OGD) has been adopted mainly by developers and data analysts, while the majority of people cannot make use of it, except by consuming it via an application or chart. I propose that in order to fully achieve all the benefits Open Data promises (e.g., more transparency, more participation and better informed and more empowered citizens, etc.), it is necessary to "democratize" Open Data by providing stakeholders better tools to not only consume, but to create and manipulate data. This idea goes in the same line as what happened with the Web2.0, where better tools (e.g, video and blogging platforms) helped people to create and share content. I will also present some of the projects I have been working on to create such types of tools for Open Data in general as well as Linked Data. Finally, I will describe the experience of regular users with some of these tools and the lessons learned from such sessions.

1 Introduction

The benefits of the Internet and the Web to a large portion of the world population are undisputed. Many people nowadays can access vast amounts of information and knowledge with a few click, something unimaginable a few decades ago. One may consider Wikipedia probably as the epitome of the information repositories on the Web, even competing with more traditional information sources[5]. Smaller websites, such as blogs, personal and institutional webpages also offer content —where the quality may differ greatly from one to another— for almost any area of human knowledge imaginable. Moreover, the so-called Web 2.0 era[10] is defined by the participation of users not only as mere spectators and consumers of information available in pages, but also as creators, reviewers and editors of content available in multiple formats, ranging from documents to video.

1.1 Participation requires better tools

One important factor in the increase of participation on the Web 2.0, that it is not always mentioned, was the availability of better tools for non-technically experts to use the Web[6][1]. These tools allowed people to create and publish information on the Web with minimal or no external help from technical-savvy users. For example, in the early 90s the

figure of the webmaster was almost mandatory for people to upload content on the Web but we have seen the decline of its role over the years (see Figure 1). Nowadays it is possible for many people to create and publish content on the Web with minimal or no intervention of technical experts[1]. Many of these tools were released as Open Source software (e.g, Wordpress, MediaWiki, Drupal) while in others cases, different companies offered these tools as a service (e.g., BlogSpot, Facebook, GoogleDocs, Flickr, YouTube). Using these tools, people who had little or no knowledge of web technologies were able to participate on the Web not only as observers, but also as authors and editors of content.

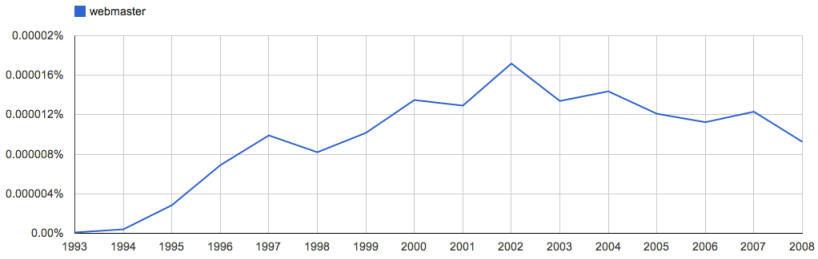

Figure 1: Occurrence of the word *webmaster* in book corpus from 1993 to 2008, via Google N-Grams.

2 Challenges of Open Data

In this sense, there is a long way to provide similar empowerment for citizens in the Open Data world. First of all, the challenges are quite different. While consuming information provided as natural text on the Web is analogous to what people has done for centuries with books and newspapers, consumption and analysis of large amounts data is something relatively new. Thus, the consumption of text on the Web is more or less straightforward, since the semiotic elements used for communicating such information are known (e.g, in western cultures, letters that composes words, which represent ideas). How to communicate data on the other side is still and open area, although initiatives to use visual[4] and auditive stimuli[3] show benefits in several disciplines[13][12].

Another aspect to consider is the context in which the data is created and used. For example, the use of mobile devices (smartphones, tablets) has become a common way to generate geographical information and publish it in social networks, such as Foursquare or Facebook. This data is later visualized in a webpage but others, but very limited options are available for users to manipulate raw data on these mobile devices. Thus, we should consider not only *what* can be done with data available on the Web to empower users, but also the context (place, medium, device) that such users will have at hand.

3 Tools for Open Data

Thus, we can see the need for better tools for users not only to consume, but to create, curate and edit data on the Web. While there are certain initiatives and companies working on

[1] It is important to note that sysadmins, devops and webmasters still play an important role maintaining websites and portals. The main difference is that the users do not need them to make every single change in their webpages.

simplifying the use of Open Data (*Socrata* and *Junar* to name a few), there is still a wide range of possibilities open in this area. Part of my work at Tetherless World Constellation has been focused on how to reduce the barriers for citizens without technical knowledge to make use of data. The final goal is to provide a set of tools that empower such users, by allowing them to create artifacts that simplify the process of consuming, exploring and sharing data on the Web. In particular, visualizations are simple artifacts that facilitates sharing and consuming data. There is evidence that visualizations can improve learning processes in corporate and educational environments[9] and it facilitate the creation, exploration and understanding of complex processes and workflows[8][7].

One of the first projects developed was *Visualbox*, a tool that facilitates the creation of visualizations based on Linked Open Data. Visualbox provides an integrated environment to retrieve the data from SPARQL endpoints and generate different types visualizations, from traditional charts to maps and dynamic graphs. In 2012 we presented a session at the Mozilla Festival[2] showing Visualbox, where participants created a series of visualizations using Linked Data obtained from *Dbpedia*[2] and *Data.gov.uk*[11]. Participants were divided in several groups that collaborated to create several maps, graphs and different types of charts showing information data about buildings, wars and musical influences. After the session, the feedback from many participants was positive by how easy was to create visualizations with our tool. Some comments referring to Visualbox: "[Visualbox] *is really easy to use, I like that you can create a visualization in two steps*" and "[Visualbox] *really helped me understand Linked Data*".

Figure 2: Map showing attractions in Berlin, where one attraction is described to be in London.

Two major challenges were mentioned by the participants. First, the creation of the right SPARQL query to retrieve the desired data is still difficult for non-Semantic Web experts. Most of the groups spent an important amount of time figuring out how to write these queries. In this sense, it is necessary to provide better, simpler mechanisms to create such queries. Visualbox provides a simple highlighting system to indicate syntactic errors, however in many cases people found it confusing in the semantics of the query. The second problem was the amount of data available as Linked Data and its quality was less than optimal in many cases. Several groups had interesting ideas to develop but they could not find the necessary data in Linked Data format. For others, the data was available via SPARQL endpoints, but it contained many errors and imprecisions (see Figure 2, showing

[2] http://0n.cl/8

a visualization created after the session).

Another suggestion made by more technical-savvy users was to allow programmers to add custom code to extend the functionalities of existing visualizations. This would provide more extensibility and would allow advanced users to create improved visualizations and mashups.

As mentioned above, one of the problems of using Visualbox was that much of the data available is not published as Linked Data. Because of that, I have been developing a second tool called OpenDataVis, whose goal is to allow non-experts to create visualizations based on Open Data published as CSV (comma-separated values), KML (Keyhole Markup Language) and other well-known format. The focus of OpenDataVis is to create visualizations that contains not only the data itself, but the provenance describing where and when the data was obtained, what processes were applied to it (e.g., data filtering, averaging) and what visual strategy has been used (e.g, a map, chart).

4 Conclusion

We can see that the need of better tools is becoming an important factor to provide capabilities for users to create, consume and share data on the Web. Giving powerful but simple tools for citizens to explore data and create data-based artifacts is key in democratizing the use of Open Data. These tools should support the whole lifecycle of data (collection/generation, curation, publication and reviewing). In order to fulfill the promises of what Open Data can bring (better transparency, better informed citizens), it is important to empower people so they can manipulate data and not being mere consumers of what others —with better technical knowledge— can show to them.

References

[1] B. Alexander. Web 2.0: A new wave of innovation for teaching and learning? *Educause review*, 41(2):32, 2006.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735, 2007.

[3] S. Barrass and G. Kramer. Using sonification. *Multimedia Systems*, 7(1):23–31, 1999.

[4] P. Fox and J. Hendler. Changing the equation on scientific data visualization. *Science*, 331(6018):705–708, 2011.

[5] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[6] D. Giustini. How web 2.0 is changing medicine. *Bmj*, 333(7582):1283–1284, 2006.

[7] P. Gordon and C. Sensen. A pilot study into the usability of a scientific workflow construction tool. *Sun Center of Excellence for Visual Genomics, Tech. Rep*, pages 874–26, 2007.

[8] T. Green and M. Petre. Usability analysis of visual programming environments: A 'cognitive dimensions' framework. *Journal of visual languages and computing*, 7(2):131–174, 1996.

[9] J. Novak. *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*. Taylor & Francis, 2010.

[10] T. OReilly. Web 2.0: compact definition. *Message posted to http://radar. oreilly. com/archives/2005/10/web_20_compact_definition. html*, 2005.

[11] J. Sheridan and J. Tennison. Linking uk government data. *BHBL+].: http://events. linkeddata. org/ldow2010/.(Cit. on p.)*, 2010.

[12] B. Shneiderman, C. Plaisant, and B. Hesse. Improving health and healthcare with interactive visualization methods. 2013.

[13] M. Watson and P. Sanderson. Sonification supports eyes-free respiratory monitoring and task time-sharing. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(3):497–517, 2004.