Current directions for Usage Analysis and the Web of Data: The diverse ecosystem of Web of Data access mechanisms

[A Message from the USEWOD Chairs]

Markus Luczak-Roesch
University of Southampton
Electronics and Computer Science
mlr1m12@soton.ac.uk

Laura Hollink
Centrum Wiskunde & Informatika
Information Access Group
l.hollink@cwi.nl

Bettina Berendt
KU Leuven
Department of Computer Science
bettina.berendt@cs.kuleuven.be

1. INTRODUCTION

Usage mining always was and still is a key topic for research in the context of the Web [16]. This is evidenced by the series of papers that appear in the scientific tracks of the WWW conference year by year. Web usage is being studied to create economic value by placing targeted ads or delivering personalized content, but also in order to better understand how people behave online in mass movements and collective action.

For more than half a decade we have been able to witness an increasing use of the Web of Data. That is the part of the Web that is not primarily meant to be consumed by human users directly through a browser interface, but instead by machines, which can create data mash-ups dynamically from various available sources. However, research on Web-of-Data usage is, today, not prominent in the main tracks of the WWW conference. Our community's understanding of methods and techniques necessary to analyze Web of Data usage logs, and the interpretation of the findings, are lagging behind when compared to the analysis of classical Web browsing or keyword search sessions of human Web users.

Since 2011, the USEWOD workshop series[1] has advanced the research agenda on usage analysis in the context of the Web of Data [1], which has undergone an interesting evolution in this period. At its inception, the workshop was closely aligned with the emerging Linked Data community effort, which today has become a widely adopted data publishing standard. Through the years that followed, USEWOD has accompanied the development of Linked Data Fragments as an alternative way to query Linked Data inspired by hypermedia principles, the charters for CSV and statistical data on the Web, and – most recently – the invention of Wikidata, a community-created knowledge base feeding structured data into Wikipedia across the boundaries of language versions.

From academic to government data, from complex SPARQL queries to Linked Data Fragments, from DBpedia to Wikidata: the data sources on the Web of Data and the ways in which these sources can be created and consumed vary greatly and raise fundamental questions. These include: (a) What kind of usage is traceable for the different access mechanisms? (b) Which actionable conclusions can we draw from observed usage patterns? (c) What are the benefits and who are the beneficiaries of usage analysis and its applications; do these go beyond advertising and personalization?

Here we give a brief report about this fundamentally new diversity of data sources and access methods, and outline a selection of current research directions for usage analysis of the diverse data sources ecosystem that the Web of Data has become.

[1] http://usewod.org

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2891068.

2. HOW TO INTERACT WITH THE WEB OF DATA?

The question how usage of Web data can be analyzed is preceded by the question how people typically interact with it, because the access and retrieval mechanisms used by data consumers impact the traces data publishers can monitor.

The most obvious differentiation to be made is the level of granularity of the data access. Data dumps, be it CSV or JSON files, Excel spreadsheets, domain-specific files, or RDF dumps of Linked Data sources, constitute the most commonly used form of raw data on the Web. Consumers download these dumps into their infrastructure, and any fine-grained interaction with resources incorporated in them happens invisibly for the data publisher. While this practice is clearly understandable from a performance point of view, the flip side is that it increases the danger that a deep Web of Data emerges. This deep Web of Data is the data mapping and linking that is hidden within the data consumers' infrastructures without being contributed back to the open Web of Data [12] – a fundamental break with the principle of the global data space the Web of Data is intended to be [6].

Inherent to the self-descriptive nature of Linked Data – that part of the Web of Data that strictly adheres to four principles that are tightly bound to the Web architecture [6] – is the possibility to consume data by hyperlink-based data discovery. This allows data publishers to trace access to raw data on the level of individual data objects. This can be used to analyze which resources of a data set are used at which frequency, and in combination with uncovering remote referrers it can even be exploited to discover new relationships [14].

While this resource-level data access allows one to analyze which data objects are consumed, it still hides which properties of the data objects are relevant for the user's task. This level of data access granularity only becomes visible to the data publisher when an interface for complex queries against the raw data is provided and opens up possibilities to determine the information needs down to

particular requirements of the schema used by the data consumers [4] or improve data pre-fetching and caching strategies for RDF repositories [7].

Complex query interfaces transfer a significant proportion of the data analysis burden (namely the costly data selection process) to the server. This has been proven problematic in terms of performance and reliability of the service endpoints [3] and has given rise to decentralised approaches to complex queries against the Web of Data, such as Linked Data Fragments [17] and Linked Data queries [5].

Resource level access, complex and Linked Data queries as well as Linked Data Fragments are specifically centered around data shared in conformance with the Linked Data principles. This does not mean that we moved into an age of multiple Webs, a Web of Data and a Web of documents, which are ultimately disjoint. A significant amount of structured data is embedded in Web pages today [2]. While this limits the opportunities for data publishers to trace the usage of that embedded data directly, it plays a crucial role in modern eCommerce when it is extracted [13] and thus may be indirectly tracked when consumers are driven to Web sites as a result of improved ranking in search engines.

Another example for this 'single Web hypothesis' is the community-curated knowledge base Wikidata [18]. Wikidata is the structured data backbone of Wikipedia, which supports the consistent handling of structured data in Wikipedia infoboxes across language versions. Operated by an extended instance of the MediaWiki software, this data source offers raw data access not only through a standard Linked Data as well as multiple complex query interfaces, but also through the Wiki interface and API. This is noteworthy as it provides feedback about the frequency of data usage and data edits to both, the data publisher and all data consumers.

3. RESEARCH DIRECTIONS

Our brief overview of Web of Data access mechanisms shows how the Web of Data landscape has significantly changed over the recent years. This increasing diversity is reflected in the data embodied by the series of USEWOD research datasets [9, 10, 11, 8], which include not only log data from widely-used Linked Data sets such as DBpedia[2], Linked Geo Data[3], Bio2RDF[4], and BioPortal[5], but also the Linked Data Fragments interface to DBpedia[6] and, most recently, unified resource access logs from Wikidata[7]. Given this broad range of Web of Data usage data the USEWOD dataset still is the reference resource for research in this space.

Alternative ways to publish usage data in a way that is more natural to the Web architecture have been proposed [15]. Such a direct link between openly shared data and its usage may ultimately benefit adaptive Web applications that rely on user feedback, for example by reformulating queries or replacing schema primitives in the adopted data model. However, these approaches to usage data publication need to be expanded to cover the same diversity of usage data as the USEWOD dataset, with particular emphasis on the read-write Web of Data. We currently also do not see that sharing usage data becomes an integral part of every open data set. This is not only a matter of technology but mostly a question of whether it would be legally and ethically correct to do this.

The papers and talks in this year's USEWOD edition highlight important aspects of the diverse ways in which the Web of Data can be accessed (as described above), and they also point towards further ways of understanding (as in studying the metadata of Web-published objects and through this study re-creating usage and re-publication trajectories).

Apart from investigating aspects of the key recent developments affecting the Web of Data and its usage analysis, the theme on the diverse ecosystem of Web of Data access mechanisms is influenced by – and will likely influence – wider issues on the Web and in Web Science. These include data management and query processing architectures, search and recommendation algorithms, human computer interaction questions, and new fields arising from these, such as what could be termed H+CI: human-and-other-intelligences computer interaction.

[2] http://dbpedia.org
[3] http://linkedgeodata.org/
[4] http://bio2rdf.org/
[5] http://bioportal.bioontology.org/
[6] http://fragments.dbpedia.org/
[7] https://www.wikidata.org/wiki/Wikidata:Data_access

References

[1] B. Berendt, L. Hollink, V. Hollink, M. Luczak-Rösch, K. Möller, and D. Vallet. Usage analysis and the web of data. *ACM SIGIR Forum*, 45(1):63–69, 2011.

[2] C. Bizer, K. Eckert, R. Meusel, H. Mühleisen, M. Schuhmacher, and J. Völker. Deployment of rdfa, microdata, and microformats on the web–a quantitative analysis. In *The semantic web–ISWC 2013*, pages 17–32. Springer, 2013.

[3] C. Buil-Aranda, A. Hogan, J. Umbrich, and P.-Y. Vandenbussche. Sparql web-querying infrastructure: Ready for action? In *The Semantic Web–ISWC 2013*, pages 277–293. Springer, 2013.

[4] K. Elbedweihy, S. Mazumdar, A. E. Cano, S. N. Wrigley, and F. Ciravegna. Identifying information needs by modelling collective query patterns. *COLD*, 782, 2011.

[5] O. Hartig and J. Pérez. Ldql: A query language for the web of linked data. In *The Semantic Web-ISWC 2015*, pages 73–91. Springer, 2015.

[6] T. Heath and C. Bizer. Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1):1–136, 2011.

[7] J. Lorey and F. Naumann. Caching and prefetching strategies for sparql queries. In *The Semantic Web: ESWC 2013 Satellite Events*, pages 46–65. Springer, 2013.

[8] M. Luczak-Roesch, S. Aljaloud, B. Berendt, and L. Hollink. USEWOD 2016 Research Dataset. Published on the University of Southampton eprints repository, http://dx.doi.org/10.5258/SOTON/385344, 2016.

[9] M. Luczak-Roesch, B. Berendt, and L. Hollink. USEWOD 2013 Research Dataset. Published on the University of Southampton eprints repository, http://dx.doi.org/10.5258/SOTON/379399, 2013.

[10] M. Luczak-Roesch, B. Berendt, and L. Hollink. USEWOD 2014 Research Dataset. Published on the University of Southampton eprints repository, http://dx.doi.org/10.5258/SOTON/379401, 2014.

[11] M. Luczak-Roesch, B. Berendt, and L. Hollink. USEWOD 2015 Research Dataset. Published on the University of Southampton eprints repository, http://dx.doi.org/10.5258/SOTON/379407, 2015.

[12] M. Luczak-Rösch, E. Simperl, S. Stadtmüller, and T. Käfer. The role of ontology engineering in linked data publishing and management: An empirical study. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 10(3):74–91, 2014.

[13] R. Meusel, A. Primpeli, C. Meilicke, H. Paulheim, and C. Bizer. Exploiting microdata annotations to consistently categorize product offers at web scale. In *E-Commerce and Web Technologies*, pages 83–99. Springer, 2015.

[14] H. Mühleisen and A. Jentzsch. Augmenting the web of data using referers. In *LDOW*. Citeseer, 2011.

[15] M. Saleem, M. I. Ali, A. Hogan, Q. Mehmood, and A.-C. N. Ngomo. Lsq: The linked sparql queries dataset. In *The Semantic Web-ISWC 2015*, pages 261–269. Springer, 2015.

[16] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, 1(2):12–23, 2000.

[17] R. Verborgh, M. Vander Sande, P. Colpaert, S. Coppens, E. Mannens, and R. Van de Walle. Web-scale querying through linked data fragments. In *7th Workshop on Linked Data on the Web*, 2014.

[18] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.