Fostering a Lively and Tenacious Web of Data

Aidan Hogan[1,*]

[1]IMFD; DCC, Universidad de Chile

Abstract

While the Web of Data continues to mature, it can still suffer from inertia (being slow to change) and impermanence (losing track of the past). Addressing these two flaws in a meaningful way requires additional work to better understand and harness dynamics on the Web of Data, not only in terms of data, but also in terms of queries, links, websites, definitions, demands, etc. In this paper, we highlight key challenges relating to dynamics on the Web of Data. We outline issues for the Web of Data that may arise if such challenges are left neglected: stale or forgotten data, incorrect results, unchecked vandalism, biased conclusions, etc. We discuss research lines to address such challenges relating to representations, modelling, prediction, etc. Within these research lines we identify key trade-offs, a better understanding of which may help us to transition towards a more lively and tenacious Web of Data that is better equipped to serve a changing world.

Keywords

Web of Data, Dynamics, RDF, Linked Data

1. Introduction

The Web of Data is a vision nearly as old as the Web itself [1], wherein data are published in structured formats that allow for increased levels of automation in everyday tasks [2]. More than two decades on, various foundational parts of the puzzle are now in place to realise a particular instance of this vision, called the Semantic Web. We have RDF [3], in various flavours, for publishing data on the Web in an interoperable way. We have RDFS [4] and OWL [5] for defining the vocabulary used in RDF data in a declarative manner, enabling automated inference. We have SPARQL [6] for evaluating complex queries over RDF datasets from across the Web. And we have languages such as SHACL [7] and ShEx [8] that allow for defining declarative constraints and validating elements of RDF data with respect to those constraints. We have also seen two major initiatives putting these standards into practice on the Web. The first initiative, Linked Data [9], has resulted in hundreds of large-scale interlinked datasets being published on the Web using the Semantic Web standards [9]. The second initiative, Embedded Metadata, has seen the RDF-based formats RDFa [10] and JSON-LD [11] being used on an estimated 44.2% and 37.7% of accessible websites, respectively,[1] using vocabularies such as that of schema.org [12].

MEPDaW'22 – Managing the Evolution and Preservation of the Data Web

*Corresponding author.

ahogan@dcc.uchile.cl (A. Hogan)

https://aidanhogan.org/ (A. Hogan)

0000-0001-9482-1982 (A. Hogan)

[1]According to https://w3techs.com/technologies/overview/structured_data, as of the time of writing (2022-12-27).

These achievements are just some of the fruits of more than two decades of planning, research, development, discussion and pondering about the Web of Data.

The vision of the Web of Data espoused several decades ago has, in a very tangible sense, already been realised to a significant extent. However, this emerging Web of Data is still in its infancy – or perhaps adolescence – and has a long (possibly infinite) way to go before becoming a fully-fledged realisation of the original vision [1]. Herein, we identify two aspects in which the current Web of Data leaves much to be desired, relating to *changes* in data and the world they describe, and to *dynamics*, which considers changes over time:

Inertia: The current Web of Data is prone to being slow to change.

Impermanence: The Web of Data is prone to losing track of the past.

While these issues are not specific to the Web of Data but rather affect the Web more generally, we will argue in what follows that the causes and effects of these two issues are distinct in the former case, and thus require dedicated research and solutions.

2. Inertia

The current Web of Data tends to lag behind the real world. While the Web itself is prone to the same issue, the causes and effects are somewhat distinct in the case of the Web of Data.

In terms of *causes*, many of the datasets published, for example, as Linked Data, were produced via an Extract–Transform–Load (ETL) process from a source dataset, producing a dump of RDF loaded into a store to enable browsing and querying. Many such datasets are irregularly refreshed from their sources [13], meaning that their RDF version may often lag behind the original sources. For example, DBpedia [14], a hugely-influential Linked Dataset extracted from Wikipedia, naturally lags behind in terms of updates to Wikipedia itself (though there are major efforts to address this that have improved this issue over time [15]). Other datasets may lag years behind their respective source; for example, LinkedGeoData [16], a widely-used RDF dataset based on OpenStreetMap, had not been updated since 2015 at the time of writing.

In terms of the *effects* of inertia, we can identify (at least) the following:

Stale/Missing/Incorrect Results: The results returned by systems that exploit such data may be out-of-date, less complete, or incorrect. When answering a complex query, applying reasoning, etc., sometimes one stale triple can dramatically influence results.

Delayed curation: If data are not directly modifiable – for example, if they are the result of an indirect ETL process – they are not easily corrected. Thus, for example, erroneous data due to vandalism [17], human error, etc., can persist for a long time.

Biased results: Queries and other forms of analyses may yield results skewed towards the past; for example, bibliometrics based on older datasets will disproportionately disfavour junior researchers who have a greater proportion of recent publications, citations, etc.

3. Impermanence

Many of the datasets once published as part of the Web of Data are no longer online. A significant portion of the datasets currently included in the LOD Cloud do not appear to be available; for example, though the domain kasabi.com is currently offline, fifteen datasets listed on the LOD Cloud point to resources on this single website. Aside from datasets, services on the Web of Data are also impermanent. The SPARQLES service [18] currently estimates that over half of the SPARQL endpoints once available and reported on DataHub are now offline, and appear permanently so.[2] Other services – such as key search engines [19, 20], data catalogues [21], etc. – are no longer available. Are these resources still available, in some fashion, on the Web? If they are not available any more, what have we lost as a result? While the Web maintains, for example, the Internet Archive, it is not clear what sort of coverage this would have for the Web of Data, and, in particular, for very large Linked Datasets.

In terms of *causes*, it is still early days for the Web of Data. In the late 00's, there was a significant amount of progress and excitement surrounding Linked Data [9], which led to many companies, organisations, governments and researchers contributing data and services to this burgeoning initiative – without necessarily having a long-term archiving or sustainability policy. When usage of these datasets was less than perhaps hoped, and hosting costs became difficult to justify, they fell into disuse and were not maintained. Many datasets were published by research groups, some of which became key resources on the Web of Data, while others were quickly abandoned, due perhaps to being student projects, or to the researchers moving on to other projects or changing affiliation, or to the datasets gaining little traction. Academic venues began accepting dataset papers [22] that – though commendable for those datasets that have lasting impact – may have incentivised publishing datasets to "get a paper", but not necessarily keeping those datasets alive beyond the paper's publication. In general, even with the best of intentions initially, keeping datasets, services, etc., available in the long term can be a costly exercise, and one that may be difficult to justify when impact appears limited.[3]

In terms of the *effects* of impermanence, we can mention:

Lost data: Datasets are at risk of being lost to time and indifference. A dataset that has not received a lot of attention thus far may become of key interest in the future. This effect is particularly worrisome for datasets that involve considerable human effort to create, or are otherwise not easily reproducible (i.e., rather than purely ETL-generated datasets).

Link rot: One cannot have a Web of Data without links, which have been a key focus of initiatives such as Linked Data. However, as datasets increasingly go offline, link rot becomes increasingly common: links that once worked now return a 404, domain not found, or similar error. Thus a dataset going offline can affect not only that dataset, but other datasets that link to it. This can be particularly problematic for datasets that use the "direct link" mechanism [2], rather than minting local IRIs for each resource linked to external datasets via `owl:sameAs`; for example, if a hypothetical dataset `ex1` holds the

[2] https://sparqles.ai.wu.ac.at/availability

[3] The current author is also plenty guilty of not being able to keep datasets, demos, services, etc., that he is involved with online, including those in published papers. Some of these involved a change of affiliation and losing access to the host machines; others involved the work of students who leave for other ventures; and so forth.

triple `ex1:PiscoSour ex1:country ex2:C123`, thus pointing to a dataset `ex2` for details about the country via a dereferenceable IRI `ex:C123`, but `ex2` later goes offline, important context is lost for `ex1`: we may not even know which country it speaks about.

4. A Lively and Tenacious Web of Data?

Much like on the current Web, the issues of inertia and impermanence can likely only ever be mitigated, not ever solved, on the Web of Data. They will continue to be issues for the foreseeable future, though hopefully less and less so, as we move towards a more *lively* Web of Data, wherein changes in the real-world, and in the underlying source datasets, are reflected in a timely manner; and a more *tenacious* Web of Data, where data, resources, etc., are not so easily lost to time. This will likely evolve over time, as experimental datasets go offline, and key datasets gain further support and resources. We highlight three directions in which the community can foster and thereafter accelerate this evolution:

Rethinking RDF production: While ETL-produced datasets have been crucial for bootstrapping Linked Data, they have also led to a high degree of inertia, as mentioned in the case of DBpedia [14]. One way to address this issue is to rather provide wrappers that can extract resources from the original source *online*, per DBpedia Live [15]. This may also address issues of impermanence, where a wrapper, rather than an explicit and potentially large dataset, may be easier to host. However, wrapper-based approaches also have potential drawbacks regarding, for example, not being able to directly query the full up-to-date dataset, etc. A seemingly better alternative is to push RDF production as close as possible to the original producer of data, where either RDF is used as the primary data management format, or, as in the case of Wikidata [23], where RDF is used as a secondary format that is actively synchronised in parallel upon updates. As an example, rather than needing an ETL-based LinkedGeoData [16], it would be ideal if OpenStreetMap published live data in RDF directly, along with appropriate interfaces. This might further address issues of impermanence, where the original source will often be a more stable resource.

Better modelling and prediction of dynamics: It is infeasible to ever locally achieve a consistent, up-to-date, complete view of a decentralised information space as massive as the Web, or the Web of Data. Designers of Web-based system must thus – be it implicitly or explicitly – make a trade-off between alignment/timeliness (how well results are aligned with the current version on the Web), coverage (how many websites are considered) and efficiency (how long requests take to answer and how many resources are used to answer them) [24].[4] Gains in this trade-off can be made through better understanding, modelling and prediction of dynamics. For example, imagine a hypothetical alerts-based system where users can register queries and receive notifications when there are updates to data on the open Web that match their query (similar to Google Alerts, but for the open Web of Data). Such a system could offer more up-to-date results more efficiently

[4] Most Web-based systems consider efficiency to be bounded or fixed, where the trade-off is then between alignment and coverage; for example, Google News covers only important news websites, but often captures updates up-to-the-minute; Google Search covers the broader Web, but may index data not updated in weeks.

over a wider range of sources if it had access to a hypothetical oracle that would indicate precisely when remote sources would change, and in what manner. In the absence of such an oracle in practice, this information can be approximated via models and predictions of dynamics [25], thus improving *liveliness*. Here *dynamics* refers – depending on the application – not only to data, but also to demand (what users are likely to request), etc.

Archiving to avoid data loss: To address *impermanence*, or conversely to improve *tenaciousness*, further research is needed on techniques for efficiently archiving RDF data, and querying such archives. Current practices to avoid data loss include publishing dumps of datasets on high-volume archiving services such as Figshare or Zenodo, but – though useful – these only support downloading the dump. Ideally we could have an archive that would enable, for example, versioned queries over the data archived; in order to be feasible, this would, in turn, require efficient techniques for indexing versioned RDF data. This is indeed an active area of research [26], but more work is needed.

5. Conclusions

Recent years have seen significant developments on the Web of Data. Based on these experiences, in this paper, we have highlighted two issues on which more work is needed: inertia and impermanence. Towards addressing these issues, we argue that it is necessary to rethink how RDF is produced and published, to conduct further research on models that can predict the dynamics of not only data but also queries (and potentially users, definitions, etc.), and to continue to develop efficient techniques for the archiving and querying of historical RDF data. These lines of research, we argue, will help to foster a more lively and tenacious Web of Data.

Acknowledgments

This work was supported by ANID — Millennium Science Initiative Program — Code ICN17_002, and FONDECYT Grant 1221926. I would also like to thank the organisers of MEPDaW 2022 for inviting me to give a keynote on this topic.

References

[1] T. Berners-Lee, Semantic Web Road map, W3C Design Issues, 1998. https://www.w3.org/DesignIssues/Semantic.html.

[2] A. Hogan, The Web of Data, Springer, 2020. doi:10.1007/978-3-030-51580-5.

[3] R. Cyganiak, D. Wood, M. Lanthaler, RDF 1.1 Concepts and Abstract Syntax, W3C Recommendation, 2014. https://www.w3.org/TR/rdf11-concepts/.

[4] D. Brickley, R. Guha, B. McBride, RDF Schema 1.1, W3C Recommendation, 2014. https://www.w3.org/TR/rdf-schema/.

[5] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, S. Rudolph, OWL 2 Web Ontology Language Primer (Second Edition), W3C Recommendation, 2012. https://www.w3.org/TR/owl2-primer/.

[6] S. Harris, A. Seaborne, E. Prud'hommeaux, SPARQL 1.1 Query Language, W3C Recommendation, 2013. https://www.w3.org/TR/sparql11-query/.
[7] H. Knublauch, D. Kontokostas, Shapes Constraint Language (SHACL), W3C Recommendation, 2017. https://www.w3.org/TR/shacl/.
[8] E. Prud'hommeaux, I. Boneva, J. E. Labra Gayo, G. Kellogg, Shape Expressions Language 2.1, W3C Final Community Group Report, 2019. http://shex.io/shex-semantics/.
[9] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space (1st Edition), volume 1 of *Synthesis Lectures on the Semantic Web: Theory and Technology*, Morgan & Claypool, 2011. Available from http://linkeddatabook.com/editions/1.0/.
[10] I. Herman, B. Adida, M. Sporny, M. Birbeck, RDFa 1.1 Primer – Third Edition – Rich Structured Data Markup for Web Documents, W3C Working Group Note, 2015. https://www.w3.org/TR/rdfa-primer/.
[11] G. Kellogg, P.-A. Champin, D. Longley, M. Sporny, M. Lanthaler, N. Lindström, JSON-LD 1.1 – A JSON-based Serialization for Linked Data, W3C Recommendation, 2020. https://www.w3.org/TR/json-ld11/.
[12] R. V. Guha, D. Brickley, S. Macbeth, Schema.org: evolution of structured data on the web, Commun. ACM 59 (2016) 44–51. doi:10.1145/2844544.
[13] T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, A. Hogan, Observing Linked Data Dynamics, in: P. Cimiano, Ó. Corcho, V. Presutti, L. Hollink, S. Rudolph (Eds.), The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings, volume 7882 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 213–227. doi:10.1007/978-3-642-38288-8_15.
[14] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, C. Bizer, DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia, Semantic Web 6 (2015) 167–195. doi:10.3233/SW-140134.
[15] M. Morsey, J. Lehmann, S. Auer, C. Stadler, S. Hellmann, DBpedia and the live extraction of structured data from Wikipedia, Program 46 (2012) 157–181. doi:10.1108/00330331211221828.
[16] C. Stadler, J. Lehmann, K. Höffner, S. Auer, LinkedGeoData: A core for a web of spatial open data, Semantic Web 3 (2012) 333–354. doi:10.3233/SW-2011-0052.
[17] S. Heindorf, M. Potthast, B. Stein, G. Engels, Vandalism Detection in Wikidata, in: S. Mukhopadhyay, C. Zhai, E. Bertino, F. Crestani, J. Mostafa, J. Tang, L. Si, X. Zhou, Y. Chang, Y. Li, P. Sondhi (Eds.), Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016, ACM, 2016, pp. 327–336. doi:10.1145/2983323.2983740.
[18] P. Vandenbussche, J. Umbrich, L. Matteis, A. Hogan, C. B. Aranda, SPARQLES: Monitoring public SPARQL endpoints, Semantic Web 8 (2017) 1049–1065. doi:10.3233/SW-170254.
[19] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, G. Tummarello, Sindice.com: a document-oriented lookup index for open linked data, Int. J. Metadata Semant. Ontologies 3 (2008) 37–52. doi:10.1504/IJMSO.2008.021204.
[20] A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres, S. Decker, Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine, J. Web Semant. 9 (2011) 365–401. doi:10.1016/j.websem.2011.06.004.

[21] W. Beek, L. Rietveld, S. Schlobach, F. van Harmelen, LOD Laundromat: Why the Semantic Web Needs Centralization (Even If We Don't Like It), IEEE Internet Comput. 20 (2016) 78–81. doi:10.1109/MIC.2016.43.

[22] A. Hogan, P. Hitzler, K. Janowicz, Linked Dataset description papers at the Semantic Web journal: A critical assessment, Semantic Web 7 (2016) 105–116. doi:10.3233/SW-160216.

[23] S. Malyshev, M. Krötzsch, L. González, J. Gonsior, A. Bielefeldt, Getting the Most Out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph, in: D. Vrandecic, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L. Kaffee, E. Simperl (Eds.), The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part II, volume 11137 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 376–394. doi:10.1007/978-3-030-00668-6_23.

[24] J. Umbrich, C. Gutierrez, A. Hogan, M. Karnstedt, J. X. Parreira, The ACE theorem for querying the web of data, in: L. Carr, A. H. F. Laender, B. F. Lóscio, I. King, M. Fontoura, D. Vrandecic, L. Aroyo, J. P. M. de Oliveira, F. Lima, E. Wilde (Eds.), 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume, International World Wide Web Conferences Steering Committee / ACM, 2013, pp. 133–134. doi:10.1145/2487788.2487852.

[25] R. Q. Dividino, T. Gottron, A. Scherp, G. Gröner, From Changes to Dynamics: Dynamics Analysis of Linked Open Data Sources, in: E. Demidova, S. Dietze, J. Szymanski, J. G. Breslin (Eds.), Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data co-located with the 11th Extended Semantic Web Conference, PROFILES@ESWC 2014, Anissaras, Crete, Greece, May 26, 2014, volume 1151 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2014.

[26] J. D. Fernández, J. Umbrich, A. Polleres, M. Knuth, Evaluating query and storage strategies for RDF archives, Semantic Web 10 (2019) 247–291. doi:10.3233/SW-180309.