Open Challenges for the Management and Preservation of Evolving Data on the Web

✉ Lars Gleim[1] and Stefan Decker[1,2]

[1] Chair of Information Systems, RWTH Aachen University, Aachen, Germany
{gleim,decker}@dbis.rwth-aachen.de
[2] Fraunhofer FIT, Sankt Augustin, Germany

Abstract. As the volume, variety, and velocity of data published on the Web continue to increase, the management, governance and preservation of these data play an increasingly important role. Data-driven decision making and algorithmic control systems rely on the persistent availability of critical information. However, to date, the free sharing, reuse and interoperability of data are hindered by a number of fundamental open challenges for the management and preservation of evolving data on the Web. In this work, we provide an overview of open challenges and recent efforts to address them. We then propose a data persistence layer for data management and preservation, paving the way for increased interoperability and compatibility.

Keywords: Sustainable Data Management · Data Preservation · Evolving Data · Survey · Citation · Persistence · Persistent Identifier · Archiving · Data Discovery · State Synchronization · Knowledge Graph · OSI Model

1 Introduction

Data sharing, reuse and integration are increasingly important for industry, research and government, as more and more data are being produced and collected, e.g. due to the influence of Internet of Things technologies. The continuous integration of data from a multitude of sources serves as foundation and enabler of significant productivity improvements, digital business models, and novel applications, e.g. in artificial intelligence (AI) [6]. As data-driven control and AI systems take increasingly autonomous decisions, the assurance of reliability, accountability and trust in the underlying evolving data base gains significant importance. As such, one of the emerging and fundamental problems of the Internet is the management and preservation of evolving data on the Web and in knowledge graphs.

In the following, we provide a concrete breakdown of the associated open challenges as identified from extensive analysis and literature review, as well as giving a concise overview of existing approaches addressing these challenges and references to in-depth surveys on the respective issues, where possible. We then propose the idea of a logical preservation and persistence layer, integrable into the classical OSI reference model to address these challenges in a uniform way.

2 Challenges for Data Management and Preservation

According to Moore [29], digital preservation fundamentally revolves around the principal assurance of infrastructure-independent data persistence, verifiability of authenticity and integrity of the preserved data, as well as that of the preservation environment and resource metadata. Based on this foundation and the analysis of relevant related work, we further decompose data management and preservation to describe the following *eight* principal corresponding challenges:

C1 *Archiving – How to archive a particular version of a resource immutably?*

Since resources on the Web typically change over time and frequently become entirely unavailable [23], saving corresponding archive copies is essential for their preservation. Following Moore's principle of *infrastructure independence* [29], an archiving mechanism should be externally transparent. As such, it should at least logically allow for later access to *independent copies* [10] of preserved states, even if more efficient storage representations, such as *change-based* or *timestamp-based* are used internally. The *compression* and *indexing* of such archives (cf. e.g., [10,31]), as well as structured *query mechanisms* for the temporal data in such archives (cf. e.g., [2,8,37,38], including continuous query evaluation, as e.g. in [39]) are further relevant subchallenges of archiving. A recent survey of current Web archiving initiatives can be found in [9], as well as a survey of RDF Archiving approaches in [31].

C2 *Citation – How to reference a particular resource variant persistently?*

As resources may change or disappear over time, being able to persistently reference a certain version or variant is essential to enable reliable data reuse, especially for data not controlled by the consumer. To address this challenge, it is common practice to employ persistent identifiers (PIDs). Hilse and Kothe [21] provide a good overview of the available systems and their respective strengths and weaknesses. As postulated by Kunze [24], persistence is however purely a matter of service of the resource provider, not a property of any specific naming syntax, and lastly depends on the commitment of any given repository to provide it. Whenever a PID exists, the identified resource should ideally also specify, i.e., wear this PID as its data identifier. Since data on the Web is however typically only identified using a non-persistent URL, relating resources and PIDs is another subchallenge of citation, as discussed in [35]. To avoid this issue, Gleim and Decker [13] recently proposed to reuse generic timestamped URLs as PIDs in combination with the time-based HTTP Memento retrieval mechanism [34].

C3 *Retrieval – How to retrieve a particular resource variant from an archive?*

Given a PID, it should be possible to resolve and retrieve the underlying resource version or variant using a standardized and open access mechanism to enable practical data inspection and reuse. While a wide variety of data retrieval and access mechanisms exist in theory, resolution of resources on the Web effectively converged towards HTTP as a common, best practice resolution and retrieval protocol. Nevertheless, to date, no explicit HTTP PID retrieval mechanism has emerged as a standard. Instead, a variety of approaches, typically based on the semantic of Cool URIs [33], are employed. Details on advanced HTTP resource resolution and retrieval can be found in [24,35,40,42].

C4 *Discovery – How to discover archives, data and available variants?*

Data discovery is a multi-dimensional challenge, incorporating (i) the discovery of resources available on the web (i.e., in need of archiving) and within a given archive, addressable e.g., using the Linked Data Platform [36], Web of Things Description [27] or classical XML Sitemaps [16], as well as (ii) the discovery of digital archives preserving certain data, (cf. e.g., [41]) and (iii) the discovery of different available variants and versions of individual resources, e.g. over time, both of which are addressable e.g. through features of the HTTP Memento protocol, as described in [42].

C5 *Synchronization – How to monitor resource changes and synchronize state?*

Especially for applications with continuous data exchange and backup, the ability to rapidly and efficiently detect resource changes, all well as subsequent state synchronization, is of critical importance. While depending on the type of data, efficient delta formats (such as *Diff* [4] for RDF or JSON Patch [7] for JSON data) are available, a simple and generic state synchronization mechanism for any type of data is needed. Due to the already mentioned dominance of the HTTP transfer protocol in the Web, corresponding synchronization mechanisms should similarly build upon HTTP if possible. Employing content-negotiation and other HTTP facilities, such as already done by the HTTP Memento protocol [34], would further enable the efficient discovery of historic resource versions though the usage of TimeMaps.

C6 *Provenance – How to track data origins and influences?*

Data provenance is, simply put, metadata information about data origins, influences and revisions. As data evolves, e.g., through modification, exchange and reuse, provenance information enables the tracking and tracing of responsible entities for specific changes, original authors and relevant modifications, as well as the assessment of data believability [32] and data quality [19]. The storage of data's provenance all the way down to the atomic level (insertions and deletions) can be very useful to backtrack the data transformation process and spotlight possible errors on different levels during the data life-cycle [30], as well as in the data itself. A summary of requirements for provenance on the Web is provided by Groth et al. [17], while the W3C PROV standard [18] serves a solid basis for its practical implementation and usage. Nevertheless, data provenance information is rarely collected to date. Therefore, a tighter integration of its collection with core business processes [14] and its semi-automatic collection [15] have been explored recently. Finally, a standardized connection between data and its provenance is frequently missing, e.g. for typical binary files shared on the Web.

C7 *Verifiability – How to assert data authenticity and integrity?*

To enable trust in the data management and preservation process, the verifiability of authenticity and integrity of the preserved data, as well as its provenance and other metadata, is another open challenge. While proposals to ensure authenticity and integrity do exist, e.g., through HTTP content signing [3,43], no technical mechanism for it is well established to date. Instead, current best practices typically rely on manual and labor-intensive data repository audits [1], a process not sufficiently well integrated into the digital infrastructure of the Web to enable automatic verification.

C8 *Sustainability – How to enable long-term data preservation at scale?*

Finally, all presented challenges are subject to the additional challenge of sustainability, to facilitate long-term data preservation at scale. Notably, this incorporates subchallenges such as achieving data redundancy through efficient data duplication, discovery, synchronization, archiving, etc. with minimal management overhead and cost. In general terms, the data management and preservation mechanism should maximize data persistence while introducing the minimal amount of associated overhead and cost possible. Recent work addressing this issue includes a large variety of approaches, ranging from decentralized, usage-based data caching [11], via abstract interoperability models [14], to full architecture proposals for sustainable publishing and querying of distributed Linked Data archives on the Web [42]. Relating back to Moore's theory of digital preservation [29], the principal assurance of infrastructure-independence finally mandates the establishment of an open ecosystem, capable of adapting to change, e.g., of the data models, metadata structure and employed protocols.

Even though we focus on these *eight* principal challenges for data management and preservation, there are a number of additional related issues worth mentioning, pertinent to the usage of evolving data: *Appraisal* (How to assess the quality of a dataset? Cf. [44,45]), *Curation* (How to integrate data, repair imperfections and ensure data quality? Cf. [12]), *Reasoning and Prediction* (How to extract and predict knowledge from evolving knowledge graphs? Cf. [26]), and *Visualization* (How to visualize data and knowledge evolution? Cf. [5]). In the following, we propose to uniformly address these challenges using an interoperability layer for data preservation.

3 The Need for Interoperable Data Preservation

While a wide range of data management and preservation systems have been proposed in the past, their practical adoption and impact have been limited. It is our understanding, that even though viable solutions for several of the mentioned challenges are readily available in principle, such as PID systems or the PROV standard [18] for data provenance, the associated effort and overhead of adoption currently render active implementation of infrastructure for data management and preservation infeasible for all but the few stakeholders, for which it is core to their business or mission. However, we believe that the success of version control and metadata management systems such as GIT [25] in the software development domain have shown, that the availability of easy to use and readily available tooling and standards leads to the adoption of corresponding systems, even by casual users. Notably, GIT serves as an underlying, interoperable data management and persistence layer, independent of the tooling employed for software development. It is our firm conviction, that a similar interoperable and easily adoptable data management and interoperability layer is needed for data on the Web, to simplify and drive the adoption of corresponding solutions.

Inspired by an early layered model for interoperability on the Web by Melnik and Decker [28], the layered architecture of Linked Data Applications by Heitmann et al. [20] and the FactDAG interoperability model [14], we propose the introduction of the *Preservation and Persistence Layer*, as illustrated in Figure 1. We position it in

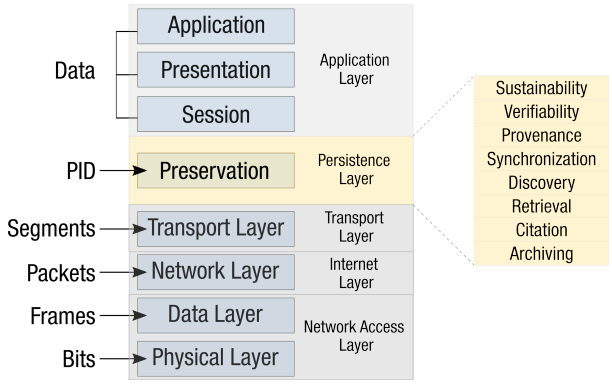

Fig. 1. Introducing a *Persistence Layer* for *Preservation* in the OSI reference model [22].

between *Transport Layer* and *Application Layer* of the TCP/IP stack, respectively in between *Transport Layer* and *Session Layer* of the classical OSI reference model [22], which characterizes and standardizes the communication functions of computing systems using abstract layers. As such, this new layer would enable uniform data persistence and preservation support for data on the Web, easily incorporable into the existing Web infrastructure and providing promising opportunities for the automation of many aspects of the preservation challenges, such as the archiving of novel resource variants, the discovery and retrieval of persistently identified archived data revisions, the collection of basic provenance information and the verification of a resource's authenticity and integrity, lastly also contributing to Web sustainability, by establishing a more uniform data management and preservation ecosystem.

4 Conclusion and Outlook

In this work, we have analyzed open challenges for the management and preservation of evolving data on the Web, identifying *eight* fundamental challenges. We have given an overview of important corresponding research addressing each of these challenges, as well as potential solution candidates. Based on these findings, we have identified the need for interoperability and deeper integration of corresponding approaches and proposed the development of a uniform *Preservation and Persistence Layer*, providing data management and preservation facilities for all types of data on the Web.

Future work should devise and evaluate means for the practical implementation and adoption of such a layer, possibly based on existing technologies such as the HTTP Memento protocol, the PROV standard for data provenance or HTTP content signing mechanisms. Implementing such a layer has the potential to uniformly and interoperably address all principal challenges of data management and preservation, furthering open data sharing, reuse and integration for applications in industry, research and government.

Acknowledgments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2023 Internet of Production – 390621612.

References

1. Ambacher, B., Ashley, K., Berry, J., Brooks, C., Dale, R.L., Flecker, D., Giaretta, D., Hamidzadeh, B., Johnson, K., Jones, M., McHugh, A., Sawyer, D., Steenbakkers, J.: Trustworthy Repositories Audit & Certification : Criteria and Checklist. Tech. rep., Center for Research Libraries (2007)
2. Anderson, J., Bendiken, A.: Transaction-time queries in Dydra. In: Joint Proceedings of MEPDaW and LDQ @ ESWC. pp. 11–19 (2016)
3. Backman, A., Richer, J., Sporny, M.: Signing HTTP Messages. Internet-Draft draft-ietf-httpbis-message-signatures-00, Internet Engineering Task Force (2020), work in Progress
4. Berners-Lee, T., Connolly, D.: Delta: an ontology for the distribution of differences between RDF graphs (2004), http://www.w3.org/DesignIssues/Diff
5. Bikakis, N., Sellis, T.: Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art. In: Proceedings of the Workshops of the EDBT/ICDT (2016)
6. KI in der Industrie 4.0: Orientierung, Anwendungsbeispiele, Handlungsempfehlungen. Tech. rep., Bundesministerium für Wirtschaft und Energie (BMWi), Berlin (2020)
7. Bryan, P.C., Nottingham, M.: JavaScript Object Notation (JSON) Patch. RFC 6902 (2013)
8. Chekol, M.W., Fionda, V., Pirrò, G.: Time Travel Queries in RDF Archives. In: Joint Proceedings of MEPDaW and LDQ @ ESWC. pp. 28–42 (2017)
9. Costa, M., Gomes, D., Silva, M.J.: The evolution of web archiving. International Journal on Digital Libraries **18**(3), 191–205 (2017)
10. Fernández, J.D., Polleres, A., Umbrich, J.: Towards Efficient Archiving of Dynamic LinkedOpen Data. In: Proceedings of DIACHRON @ ESWC. pp. 34–49 (2015)
11. Folz, P., Skaf-Molli, H., Molli, P.: CyCLaDEs: A Decentralized Cache for Triple Pattern Fragments. In: ESWC 2016: The Semantic Web. Latest Advances and New Domains, pp. 455–469. Springer International Publishing (2016)
12. Freitas, A., Curry, E.: Big Data Curation. In: New Horizons for a Data-Driven Economy, pp. 87–118. Springer International Publishing (2016)
13. Gleim, L., Decker, S.: Timestamped URLs as Persistent Identifiers. In: MEPDaW @ ISWC (2020)
14. Gleim, L., Pennekamp, J., Liebenberg, M., Buchsbaum, M., Niemietz, P., Knape, S., Epple, A., Storms, S., Trauth, D., Bergs, T., Brecher, C., Decker, S., Lakemeyer, G., Wehrle, K.: FactDAG: Formalizing Data Interoperability in an Internet of Production. IEEE Internet of Things Journal **7**(4), 3243–3253 (2020)
15. Gleim, L., Tirpitz, L., Pennekamp, J., Decker, S.: Expressing FactDAG Provenance with PROV-O. In: MEPDaW @ ISWC (2020)
16. Google Inc., Yahoo Inc., Microsoft Corporation: Sitemaps XML format (2008), https://www.sitemaps.org/protocol.html
17. Groth, P., Gil, Y., Cheney, J., Miles, S.: Requirements for Provenance on the Web. International Journal of Digital Curation **7**(1), 39–56 (2012)
18. Groth, P., Moreau, L.: PROV-Overview. W3C Working Group Note (2013)
19. Hartig, O., Zhao, J.: Using Web Data Provenance for Quality Assessment. In: Proceedings of SWPM @ ISWC. vol. 526 (2009)
20. Heitmann, B., Cyganiak, R., Hayes, C., Decker, S.: Architecture of Linked Data Applications. In: Harth, A., Hose, K., Schenkel, R. (eds.) Linked Data Manag. Princ. Tech., pp. 69–91. Chapman and Hall/CRC, 1st edn. (2014)
21. Hilse, H.W., Kothe, J.: Implementing Persistent Identifiers (2006)
22. OSI/IEC 7498-1:1994: Information technology – Open Systems Interconnection – Basic Reference Model: The Basic Model. ISO (1994)

23. Käfer, T., Abdelrahman, A., Umbrich, J., O'Byrne, P., Hogan, A.: Observing Linked Data Dynamics. In: ESWC 2013. vol. LNCS 7882, pp. 213–227 (2013)
24. Kunze, J.A.: Towards Electronic Persistence Using ARK Identifiers (2003)
25. Loeliger, J., McCullough, M.: Version Control with Git: Powerful tools and techniques for collaborative software development. O'Reilly Media (2012)
26. Margara, A., Urbani, J., van Harmelen, F., Bal, H.: Streaming the Web: Reasoning over dynamic data. Journal of Web Semantics **25**(March), 24–44 (2014)
27. Mathew, S.S., Atif, Y., Sheng, Q.Z., Maamar, Z.: Web of things: Description, discovery and integration. In: ICIOT and CPSCom. pp. 9–15. IEEE (2011)
28. Melnik, S., Decker, S.: A Layered Approach to Information Modeling and Interoperability on the Web. Proceedings of Workshop on the Semantic Web @ ECDL (2000)
29. Moore, R.: Towards a Theory of Digital Preservation. International Journal of Digital Curation **3**(1), 63–75 (2008)
30. Ngomo, A.C.N., Auer, S., Lehmann, J., Zaveri, A.: Introduction to Linked Data and Its Lifecycle on the Web. In: Reasoning Web. Reasoning on the Web in the Big Data Era, pp. 1–99. Springer International Publishing (2014)
31. Pelgrin Olivier, Galárraga Luis, Hose Katja: Towards Fully-fledged Archiving for RDF Datasets. Semantic Web Journal **1**(0), 1–20 (2020)
32. Prat, N., Madnick, S.: Measuring Data Believability: A Provenance Approach. In: HICSS. pp. 393–393. IEEE (2008)
33. Sauermann, L., Cyganiak, R., Völkel, M.: Cool URIs for the Semantic Web (2007), http://www.w3.org/TR/cooluris/
34. Van de Sompel, H., Nelson, M., Sanderson, R.: HTTP Framework for Time-Based Access to Resource States – Memento. RFC 7089 (2013)
35. Van de Sompel, H., Sanderson, R., Shankar, H., Klein, M.: Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping. International Journal of Digital Curation **9**(1), 331–342 (2014)
36. Speicher, S., Arwe, J., Malhotra, A.: Linked Data Platform 1.0. W3C Recommendation, February **26** (2015)
37. Taelman, R., Sande, M.V., Verborgh, R., Mannens, E.: Versioned Triple Pattern Fragments: A Low-cost Linked Data Interface Feature for Web Archives. Joint Proceedings of MEPDaW and LDQ @ ESWC (2017)
38. Taelman, R., Takeda, H., Sande, M.V., Verborgh, R.: The Fundamentals of Semantic Versioned Querying. In: Proceedings of SSWS @ ISWC. pp. 1–14 (2018)
39. Taelman, R., Verborgh, R., Colpaert, P., Mannens, E.: Continuous Client-Side Query Evaluation over Dynamic Linked Data. In: The Semantic Web, pp. 273–289. Springer International Publishing (2016)
40. Thompson, H.S., Orchard, D.: URNs, Namespaces and Registries (2006), https://www.w3.org/2001/tag/doc/URNsAndRegistries-50
41. Vander Sande, M., Verborgh, R., Dimou, A., Colpaert, P., Mannens, E.: Hypermedia-Based Discovery for Source Selection Using Low-Cost Linked Data Interfaces. International Journal on Semantic Web and Information Systems **12**(3), 79–110 (2016)
42. Vander Sande, M., Verborgh, R., Hochstenbach, P., Van de Sompel, H.: Toward sustainable publishing and querying of distributed Linked Data archives. Journal of Documentation **74**(1), 195–222 (2018)
43. Yasskin, J.: Signed HTTP Exchanges. Internet-Draft draft-yasskin-http-origin-signed-responses-09, Internet Engineering Task Force (Jul 2020), work in Progress
44. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for Linked Data: A Survey. Semantic Web Journal **7**(1), 63–93 (2016)
45. Zhu, H., Madnick, S., Lee, Y., Wang, R.: Data and Information Quality Research: Its Evolution and Future. In: Computing Handbook, 3rd ed. (2014)