A Decentralized Provenance Network for Linked Open Data

Fabian Kirstein[0000−0002−9064−2546]

Weizenbaum Institute for the Networked Society, Berlin, Germany
Fraunhofer FOKUS, Berlin, Germany
fabian.kirstein@fokus.fraunhofer.de

Abstract. With the growing availability of Linked Open Data (LOD) and the consequential generation of derived and aggregated data, the need for trustworthy, reproducible and accessible provenance information has increased. Yet, no consistent mechanism has been established to manage provenance data of LOD on a global dataset-level. Decentralized networks and peer-to-peer mechanisms have made their revival in the last years with blockchain and similar distributed ledger technologies. We propose a novel approach to track and store provenance information for LOD on a dataset-level by sharing an immutable, common state between data providers. The basic architecture will not disrupt existing methodologies and standards for publishing LOD, but will be transparently integrated into existing ecosystems as an additional layer to foster broad acceptance. We will investigate the application of emerging blockchain technologies and established Linked Data specifications for building this decentralized anchor of truth. We are actively involved in the design and implementation of LOD and Open Data platforms and will evaluate our approach in real-world scenarios regarding feasibility, governance, scalability and usability.

Keywords: Provenance · Distributed Ledger · Blockchain · Open Data.

1 Problem Statement

The Linked Open Data (LOD) movement is a global phenomenon, driven by the fact that additional value is generated by interlinking structured data. LOD is Linked Data, which can be distributed by everyone anytime without any restrictions. An active community has evolved around the publication and generation of LOD. A popular publisher is WikiData, which freely offers comprehensive data, completely serialized in the Resource Description Framework (RDF). [25] Other issuers release only metadata as LOD, referencing and describing in detail the actual data inventory, which usually consists of a variety of data formats. Our work focuses on the latter approach, which is typically applied by Open Data portals, aggregating public data, published by public administrations or research organizations. Well-known examples are OpenAIRE [12] for research data and the European Data Portal (EDP) [6] for data of public authorities in

Europe. In fact, many more publishers of LOD exist: The Linked Open Data Cloud[1] lists more than 1300 datasets.
The widespread distribution and availability of Open Data leads to the creation and publication of derivative or edited datasets. Data from different sources and origins is copied, aggregated, converted and/or enriched. Furthermore, claims and conclusions are inferred from (combinations of) these datasets. The traceability and repeatability of such data and its processing is critical for maintaining trust and accountability. A central foundation for that is the presence of expressive and valid provenance information for each dataset in an LOD processing chain. **No unified mechanism is established to record and track the provenance of LOD on a dataset-level.** The very nature of LOD causes barriers in establishing appropriate measures. The intrinsic reason for this claim is: LOD ecosystems constitute highly distributed and decentralized systems, where data is acquired and aggregated across distinct organizational and technical levels. An illustrating example for this is the harvesting process of LOD published by public services. Typically, harvesting is conducted in a bottom-up form, where municipal data providers publish data independently. This is followed by an aggregation towards the next higher organizational level and so forth, e.g., towards data portals of cities, federal states, etc. [9] Users and processors may fetch the data at any point in the hierarchy. A similar process is applied for scientific publications, where data is published by individual research organizations and aggregated in central hubs for scientific publications [12].
This methodology and the design of LOD leads to major challenges with respect to tracking the provenance of a single dataset: Firstly, there is no established approach for uniquely and persistently identifying a distinct dataset. Although the Linked Data principles require the use of Uniform Resource Identifiers (URIs) as unique identifiers, they can be easily reassigned and the DNS itself is transient. Especially in the domain of research data, the application of Digital Object Identifiers (DOIs) as a centralized workaround is established. Secondly, provenance information is often not set, fragmentary or not correctly forwarded in an acquisition and processing chain. Expressive and rich specifications for encoding provenance are available, foremost the W3C PROV [13]. Still, its successful adoption requires correct handling of it by each participant. Additionally, provenance information represented as plain metadata allows tampering and manipulation by malicious partners.
These limitations could be solved by establishing and agreeing on a central management system for provenance and identifier information. However, the very character of LOD is decentralization and sovereignty, involving multiple stakeholders and heterogeneous infrastructures. This environment makes the successful implementation of a centralized system infeasible. This leads to the essential hypotheses of this dissertation: **An additional, immutable and decentralized network can help to overcome current drawbacks in LOD provenance tracking and incentivize its broad application.** The recent developments in blockchain systems and related peer-to-peer (P2P) technologies will

[1] `https://www.lod-cloud.net`

be an important foundation for implementing such a network.
This approach can accomplish both: The management of provenance information through a homogeneous system and the protection of the independence of the data providers. This will support the fact that from an organizational point of view, LOD forms a highly decentralized system, which requires a single point of truth in order to ensure integrity and trust.

2 Relevancy

Due to the ongoing worldwide digitization, data has become a most valid asset and the basis of many value-added processes and business models. Although our work focuses on LOD, this is true for both, public domain data and proprietary data. The relevance of trustworthy information about the provenance and lineage of data will continue to increase. Simmhan et al. write "With a growing number of datasets available in the public domain beyond the confines of a single organization, it has become increasingly important to determine the veracity and quality of these datasets." [20] As of today, more than 2600 Open Data portals exist in the world. [16] Although they do not all serve LOD, it shows a clear tendency of increasing significance in the domain. This data is used, processed, aggregated and re-published by multiple user groups, e.g., journalists, scientists, businesses, citizens, etc. These groups will benefit highly from improved provenance information, since it will enable reproducibility and increase trust. This "proof of origin" will improve the overall quality of the data for the data consumers. This is especially true for the research community, where traceability is an ethical and legal requirement. With regard to the Open Science movement and the increasing publication of raw scientific data, provenance information will become essential. Within the LOD community, efforts for harmonization are intrinsic and serve the idea of a global interlinked knowledge graph. Well-known examples are the Linked Open Vocabularies project [24] and the Linked Data Platform specification [27]. Integrating a trustable, decentralized provenance mechanism can strengthen Linked Data as core layer for the growing data economy, not limited to LOD, and broaden its adoption. After all, the Semantic Web Stack is missing a trust layer, where provenance will be one essential building block.

3 Related Work

A lot of research was conducted in the relevant fields of our work. Our approach crosses established research of provenance for LOD with blockchain and distributed ledger technologies, which has been already examined to some degree.
In general, data provenance has been widely studied with respect to its use, subject, representation, storing and dissemination and a variety of software solutions have been developed for managing provenance. These approaches mainly focus on local data, typically generated by a particular scientific domain, e.g., Physics, Earth Sciences, etc. [20] An extensive literature review and overview

of provenance on the Web, including the Semantic Web, was published by Luc Moreau. [14]

3.1 Provenance for Linked (Open) Data

Early work on provenance for Linked Data focused on modelling RDF vocabularies and ontologies, which can be used to describe the provenance of published RDF data and query it respectively. [7] Since then, the W3C has developed PROV, a set of specifications and data models for publishing provenance information. It is widely established as interchange format for provenance data. The standard is not limited to Linked Data, but offers multiple serializations, including an OWL ontology. [13] Extensive research was conducted for effectively attaching provenance information to RDF. Common approaches include the concept of annotated RDF [23], where each triple is associated with metadata. Wylot et al. introduced a high-performance triplestore, allowing to store provenance-enriched RDF and executing queries, including close-grained provenance information. [31] Little work exists on making provenance information centrally and globally available. ProvStore is such a central service, allowing to store and publish provenance information of data, based on the PROV standard. [8] No approach exists in managing provenance information in a globally shared state.

3.2 Linked Data and Distributed Ledgers

First research exists on the connection of distributed ledgers/blockchain technologies and Linked Data/Semantic Web, spanning multiple aspects. English et al. endorse the notion of improving the persistent identification of RDF resources with blockchain. [5] Third et al. investigate several stages of extension of storing Linked Data in a distributed ledger, from a simple verification layer to a pure storage layer. [21] The InterPlanetary Linked Data (IPLD) project follows a disruptive approach, by completely lifting the data management to a decentralized network. IPLD offers a custom data structure, which is globally addressable and supports interlinking. [17] Sicilia et al. propose an immutable, decentralized storage for raw LOD based on the P2P System Interplanetary File System (IPFS) to overcome issues of availability. [19] An opposed approach makes decentralized data on the Ethereum blockchain available via Semantic Web technology, by mapping the blockchain data structures to Linked Data. [22] Applying a distributed ledger as an additional layer for provenance tracking in the domain of LOD was not proposed yet.

3.3 Blockchain and Beyond

Blockchain and related technologies are vivid topics of research, where most work focuses on privacy and security aspects. [33] The most defining and relevant work is the P2P cash system Bitcoin. [15] However, many different areas

of application have evolved. A general indicator for applying a blockchain is the presence of a decentralized environment, with multiple (untrusted) participants and the need for transparency. [32] Some is related to our proposed approach, but set in different domains with other emphasises. Rohrer et al. propose a blockchain-based system for decentralized and transparent storing of citation and reference provenance for journalistic articles on the Web. [18] Liang et al. implemented an additional provenance layer based on a blockchain network for the open source cloud solution ownCloud, which tracks every file transaction with only little overhead. [11] Other relevant work includes the vibrant ecosystem of open source blockchain projects. Ethereum is a multi-purpose, decentralized and transaction-based state machine. It includes smart contract functionality and allows to build private or public blockchain systems. [30] Hyperledger Fabric enables the creation of permissioned blockchains based on general-purpose programming languages and custom consensus mechanisms. [2] Finally, a lot of up-to-date research is conducted regarding consensus protocols for enduring Byzantine failures and ensuring a unique and correct state of a network. Cachin et al. give a comprehensive overview on the recent developments. [3]

4 Research Questions

Based on the problem statement, the related work and the recent impact of distributed ledger technologies, new approaches for addressing provenance of LOD will emerge. Our work will focus on an additional, decentralized layer, accompanying existing solutions for publishing LOD. Therefore, we formulate the following research questions, where **RQ1** represents the overall question.

RQ1: Can we manage the provision, management and traceability of provenance information for LOD datasets by applying an additional, decentralized layer?

RQ2: How can we persistently identify and represent provenance information of LOD in a globally unique way?

RQ3: What consensus and governance mechanisms can be applied to ensure the integrity of such a system?

RQ4: Which paradigms and tools are suitable to implement the proposed approach, considering expectations in flexibility, scalability and usability?

5 Hypotheses

The following hypotheses relate to the aforesaid research questions. **H1** depicts the overall hypothesis of the proposed thesis.

H1: A decentralized network, which holds a globally shared state for all data providers will improve the tracking and storing of provenance information of LOD in comparison to locally published provenance data.

H2: An immutable and transparent global database will improve the persistent and unique identification and management of provenance information over established transient approaches and enables a long-term preservation.

H3: An authority-based governance model and voting-based consensus mechanism will ensure a consistent state of the network and prevent misuse.

H4: Blockchain and related technologies can serve as a technical foundation for the proposed decentralized network.

6 Preliminary Results

In this section, we present first results and experiences from previous and ongoing work in the LOD and distributed ledger domains.
Our work on the EDP [9] has given us valuable insights into the process of LOD acquisition, processing and re-publishing. We collect Linked Data from more than 70 data publishers, in total more than 800.000 datasets. The data publishers themselves gather the data from lower organizational levels. It has been proven extremely difficult to uniquely identify a dataset in this ecosystem and to track its provenance. The required metadata simply does not exist or is incomplete. In addition, close communication with the data publisher has shown that there is an aspiration for autonomy and sovereignty. Rapid changes in existing methodologies and technologies are not endorsed. We came to the conclusion that an additional and simple to integrate solution has higher chances for adoption.
In our project Policy Compass[2], we developed a platform for mixing, extending, interpreting and visualizing Open Data. A use case is the assessment of outcome and impact of governmental policies through analyzing public available data. [10] We integrated several data sources, e.g. the EDP, Eurostat[3] and DBpedia[4]. The project has shown us a clear need for traceability and reproducibility, especially in the domain of policy evaluation and derived recommended actions. We implemented basic traceability support, by linking datasets to its original source and indicating the local provenance in derived assets. However, due to the heterogeneity of the data sources, the implementation of a more general and global provenance mechanism has proven unfeasible.
We have conducted several practical case studies with blockchain and distributed ledger technologies to classify their opportunities and challenges. Based on the public Ethereum blockchain we have implemented a decentralized digital identity management system. It allows human users to acquire a persistent identifier and link public properties, like date of birth, to it. The work is based on Ethereum smart contracts and the Decentralized Identifiers (DIDs) specification. [29] Furthermore, we used the permissioned blockchain infrastructure Hyperledger Fabric [2] for implementing a track and trace system for physical assets. It demonstrates how a decentralized network can enable data sharing and

[2] `https://policycompass.eu`
[3] `https://ec.europa.eu/eurostat`
[4] `https://wiki.dbpedia.org`

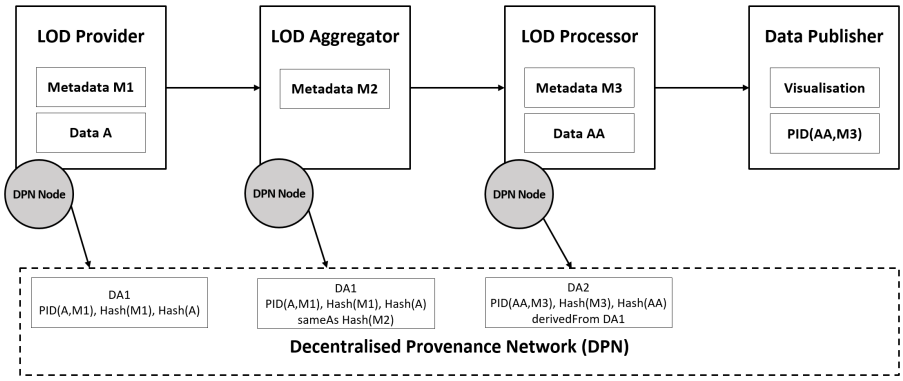

Fig. 1. An exemplary high-level process including the provenance network

cooperation beyond organizational borders. We acquired fundamental knowledge about governance, scalability and representing custom data in such a decentralized environment.

7 Approach

The overall approach is divided into four steps. Steps 1 and 2 refer to **RQ2**/**H2**, step 3 to **RQ3**/**H3** and step 4 to **RQ4**/**H4**. The overall outcome relates to **RQ1**/**H1**.

LOD and the Semantic Web follow a decentralized methodology, still some aspects require a central authority to be most effective and accurate. An indicator for that are various approaches to harmonize LOD by some kind of central stewardship, e.g., Linked Open Vocabularies (LOV) or GeoNames.org.[5] Our approach is the establishment of a decentralized network, which holds a globally shared state for all data providers and acts as an anchor of truth about provenance information. It is a single point of access for this global information, which simplifies management and traceability. Each participant of the ecosystem will act as distributed database node. The central premise is not to disrupt existing methodologies and standards, but to transparently integrate into existing ecosystems. Fig. 1 illustrates an exemplary high-level process for LOD aggregation, processing and use, including the decentralized provenance network. An LOD provider holds metadata and data and publishes a reference of them to the network. (here illustrated as hash values and a persistent identifier) An LOD aggregator copies only the metadata and extends the original reference accordingly. (indicated as `sameAs`) An LOD processor creates new data based on the original data and adds a new reference to the network, including a `derivedFrom` indication. Finally, a data publisher creates a visualization of the data, linking it to the reference in the network, allowing a clear tracking of the provenance of the data.

[5] `https://www.geonames.org`

1. Formal definition of a dataset: LOD constitutes a set of triples (aka statements), forming a multigraph. Our work does not base upon this smallest entity of LOD. Typically, a distinct subset of triples forms a self-contained information unit, restricted by pre-defined boundaries. Concepts like named graphs reflect this approach. [4] In the first step we will derive a formal definition of what can be considered a distinct *dataset*. This includes a URI schema, graph constraints and publication guidelines. We contemplate to base it on well-established standards and best practices. The Data Catalogue Vocabulary (DCAT) [26] will act as a principle recommendation for describing the metadata. The Linked Data Platform (LDP) specification [27] offers a reference for publishing the datasets on the Web. A validation mechanism will be based on the Shapes Constraint Language (SHACL) [28]. The result of this step will be a practical tool set to publish valid datasets and the groundwork for the following steps.

2. Definition of the identifier and provenance data models: Published datasets from a provider can be considered *local*, since they are initially confined to the providers network and are addressed via a transient URL. The proposed decentralized network forms a *global* context, since it is shared by many data providers and leverages all available datasets. In this step, we will essentially model and define the mapping from a local to a global context. This includes two aspects: Firstly, the global representation of a persistent identifier and its linking to the actual local dataset will be modelled. It is important here to consider changes and relocations of the local identifier and to provide the means to perform a mapping of the identifiers multi-directionally. Existing decentralized identifier concepts will act as guidance here. Secondly, the actual global provenance data model will be designed. Based on the global identifiers, we will provide a compact and basic model to represent the provenance of a dataset. It will utilize a subset of the methodology and ontology of W3C PROV and its core concepts `Entity`, `Agent` and `Activity`. [13] The outcome of this will be a comprehensive specification of the data models, alongside a proof-of-concept implementation.

3. Design of the decentralized network methodology: In this step, we will design the fundamental architecture of the decentralized provenance network, essentially regarding agent management, governance model, security aspects and consensus mechanisms. Essentially, a change in the globally shared state needs corporative validation and confirmation of the network. Only thereby the correctness and integrity can be guaranteed. We think that an authority-based governance model and voting-based consensus mechanism will ensure a consistent state of the network. Every LOD data provider will have a verifiable identity, authorizing them as a valid member of the network. This authority will be granted by a proof of ownership, e.g., of a local LOD endpoint. A state change of the network can be issued by each participant, but requires approval by the majority of the other authorized participants. Hence, a voting is performed, ensuring that not a single participant can publish defective or wrong data. Eventually, the transparency of the decentralized network will offer an additional layer of

governance. It enables an open and immediate quality assessment and increases the barrier for publishing faulty information.
We will evaluate these assumptions against real-world LOD ecosystems and publication schemes. The outcome will lead to accurate guidelines about *who* will be allowed to add *what* data *when* in the shared store, formed by the network. Especially, the on-boarding process in this decentralized environment needs to be investigated. These assumptions will be evaluated with practical artifacts, either based on existing technologies or individually implemented.

4. Implementation and evaluation of the provenance network: In the final step, we will implement the network and apply it in a production environment. With blockchain and similar distributed ledger technologies, decentralized networks and peer-to-peer mechanisms have made their revival in the last years. A variety of tools offer improved possibilities for sharing a common state and reaching consensus in a decentralized environment. Multiple implementations exist for building customized decentralized networks with desired characteristics: from public, permissionless to private, permissioned networks, including custom security and consensus protocols. These recent developments can operate as a technical foundation for the proposed decentralized network. Yet, your work will not be limited to blockchain and distributed ledger technologies, but will also consider traditional peer-to-peer mechanisms and implementations.
The work here will be mainly conducted on two levels. (1) Providing the means for actually creating the network. This includes a deployable node and a proper on-boarding process to become a participant in the network. The setup of a node is envisioned to be as straight-forward as possible. Container technologies, like Docker, might be suitable approaches here. [1] We will put an emphasis on scalability and performance and take into account typical data volumes and throughputs of LOD systems. (2) Create an approach and implementation for effectively interacting with the network. A straight-forward and easy integration into existing LOD publication concepts is desired here. The most native method here constitutes SPARQL. We think that the least disruptive integration approach would be a proxy for a standard SPARQL endpoint, allowing users to annotate publication queries with provenance information. The proxy will extract these annotations, process them and trigger a change of state in the network, when necessary. It has to be noted that the operators of (1) and (2) can be disjoint, so not every data provider has to provide a node and vice versa. The outcome of this step will be a fully working prototype.

8 Evaluation Plan

We plan to evaluate our hypotheses with the following four approaches.

1. Working prototype: Based on a proof-of-concept system, we will test and evaluate the fundamental functionality of our approach. Test data will be generated in real-world volumes. Synthetic, but representative stakeholders and actors will

use the network. We will use the results and findings for improving our approach in an iterative manner.

2. Application in a production environment: We are actively involved in the implementation of LOD portals, like the EDP. Hence, we will apply our solution in a production environment and monitor its qualities and possible adoption. A cooperation with external stakeholders, like original data publishers and data users are requested.

3. Practical usefulness: We will measure and qualify multiple characteristics of the synthetic and the production system. This includes overall performance, throughput, maximum load and scalability. Since no system for comparison exits, we will evaluate the findings on established expectations for central solutions, especially for provenance tracking.

4. User studies: The rate of adoption of such a system, is highly dependent on user acceptance. We will conduct user studies within two different user groups: (1) Data providers will be asked to join the network by integrating it into their systems. (2) Data consumers will use the provided information to express provenance statements about given datasets. It is planned to conduct the user studies twice, with a working prototype and a production version.

9 Reflections

To the best of our knowledge, the proposed research questions and the proposed approach is a novelty. There does not exist an established solution for a tamper-proof and globally accessible ledger for provenance information about LOD. The recent developments and successful real-world applications of blockchain and similar networks have demonstrated the success and acceptance of a globally shared state-machine. However, we think that blockchain still has a long way to go and are aware of its current limitations. A complete migration from established centralized systems and architectures, especially in LOD, is improbable. An additional, decentralized layer, respecting established mechanisms and standards will have a much better chance for adoption. We are actively involved in many production LOD, Open Data and Open Science projects. Among others, this includes the development of the EDP and the design and installation of a research data platform for the Weizenbaum Institute for the Networked Society. This allows us to work closely with many relevant stakeholders and consider their needs and requirements, e.g., the data providers, users or system administrators.

10 Acknowledgements

This work has been funded by the Federal Ministry of Education and Research of Germany (BMBF) under grant no. 16DII111 ("Deutsches Internet-Institut") and is supervised by Prof. Manfred Hauswirth.

References

1. Anderson, C.: Docker [Software engineering]. IEEE Software **32**(3), 102–c3 (May 2015). https://doi.org/10.1109/MS.2015.62
2. Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., Enyeart, D., Ferris, C., Laventman, G., Manevich, Y., Muralidharan, S., Murthy, C., Nguyen, B., Sethi, M., Singh, G., Smith, K., Sorniotti, A., Stathakopoulou, C., Vukolić, M., Cocco, S.W., Yellick, J.: Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains. Proceedings of the Thirteenth EuroSys Conference on - EuroSys '18 pp. 1–15 (2018). https://doi.org/10.1145/3190508.3190538
3. Cachin, C., Vukolić, M.: Blockchain Consensus Protocols in the Wild. arXiv:1707.01873 [cs] (Jul 2017)
4. Carroll, J.J., Bizer, C., Hayes, P.J., Stickler, P.: Named graphs. Journal of Web Semantics **3**, 247–267 (2005). https://doi.org/10.1016/j.websem.2005.09.001
5. English, M., Auer, S., Domingue, J.: Block chain technologies & the semantic web: A framework for symbiotic development. In: Computer Science Conference for University of Bonn Students, J. Lehmann, H. Thakkar, L. Halilaj, and R. Asmat, Eds. pp. 47–61 (2016)
6. European Data Portal: The European Data Portal: Opening up Europe's public data, `https://www.europeandataportal.eu/sites/default/files/edp_factsheet_what_is_edp_project_online.pdf`, (Accessed: 12.04.2019)
7. Hartig, O., Zhao, J.: Publishing and Consuming Provenance Metadata on the Web of Linked Data. In: McGuinness, D.L., Michaelis, J.R., Moreau, L. (eds.) Provenance and Annotation of Data and Processes. pp. 78–90. Lecture Notes in Computer Science, Springer Berlin Heidelberg (2010)
8. Huynh, T.D., Moreau, L.: ProvStore: A Public Provenance Repository. In: Ludäscher, B., Plale, B. (eds.) Provenance and Annotation of Data and Processes. pp. 275–277. Lecture Notes in Computer Science, Springer International Publishing (2015)
9. Kirstein, F., Dittwald, B., Dutkowski, S., Glikman, Y., Schimmler, S., Hauswirth, M.: Linked Data in the European Data Portal: A Comprehensive Platform for Applying DCAT-AP. In: EGOV2019 – Joint Conference EGOV-CeDEM-EPART 2019 (2019)
10. Kokkinakos, P., Koutras, C., Markaki, O., Koussouris, S., Trutnev, D., Glikman, Y.: Assessing Governmental Policies' Impact Through Prosperity Indicators and Open Data. In: Proceedings of the 2014 Conference on Electronic Governance and Open Society: Challenges in Eurasia. pp. 70–74. EGOSE '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2729104.2729134
11. Liang, X., Shetty, S., Tosh, D., Kamhoua, C., Kwiat, K., Njilla, L.: ProvChain: A Blockchain-Based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability. In: 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID). pp. 468–477 (May 2017). https://doi.org/10.1109/CCGRID.2017.8
12. Manghi, P., Manola, N., Horstmann, W., Peters, D.: An Infrastructure for Managing EC Funded Research Output: The OpenAIRE Project. The Grey Journal (TGJ) : An International Journal on Grey Literature **6**(1), 31–39 (2010)
13. Missier, P., Belhajjame, K., Cheney, J.: The W3C PROV Family of Specifications for Modelling Provenance Metadata. In: Proceedings of the 16th International Conference on Extending Database Technology. pp. 773–776. EDBT '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2452376.2452478

14. Moreau, L.: The Foundations for Provenance on the Web. Foundations and Trends in Web Science **2**, 99–241 (Nov 2010)
15. Nakamoto, S., et al.: Bitcoin: A peer-to-peer electronic cash system (2008)
16. OpenDataSoft: A Comprehensive List of 2600+ Open Data Portals around the World, `https://www.opendatasoft.com/a-comprehensive-list-of-all-open-data-portals-around-the-world/`, (Accessed: 11.04.2019)
17. Protocol Labs: IPLD - The Data Model of the Content-Addressable Web, `https://ipld.io/`, (Accessed: 15.04.2019)
18. Rohrer, E., Heidel, S., Tschorsch, F.: Webchain: Verifiable Citations and References for the World Wide Web . https://doi.org/10.14279/depositonce-8376
19. Sicilia, M.A., Sánchez-Alonso, S., García-Barriocanal, E.: Sharing Linked Open Data over Peer-to-Peer Distributed File Systems: The Case of IPFS. In: Research Conference on Metadata and Semantics Research. pp. 3–14. Springer (2016)
20. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance techniques. No. IUB-CS-TR618. (September 2005) p. 25 (2005)
21. Third, A., Domingue, J.: LinkChains: Exploring the Space of Decentralised Trustworthy Linked Data. DeSemWeb@ISWC (2017)
22. Third, A., Domingue, J.: Linked Data Indexing of Distributed Ledgers. In: Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion. pp. 1431–1436. ACM Press, Perth, Australia (2017). https://doi.org/10.1145/3041021.3053895
23. Udrea, O., Recupero, D.R., Subrahmanian, V.S.: Annotated RDF. ACM Trans. Comput. Logic **11**(2), 10:1–10:41 (Jan 2010). https://doi.org/10.1145/1656242.1656245
24. Vandenbussche, P.Y., Atemezing, G.A., Poveda-Villalón, M., Vatant, B.: Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web. Semantic Web **8**(3), 437–452 (2017)
25. Vrandečić, D., Krötzsch, M.: Wikidata: A Free Collaborative Knowledgebase. Commun. ACM **57**(10), 78–85 (Sep 2014). https://doi.org/10.1145/2629489
26. W3C: Data Catalog Vocabulary (DCAT), `https://www.w3.org/TR/vocab-dcat/`
27. W3C: Linked Data Platform 1.0, `https://www.w3.org/TR/ldp/`
28. W3C: Shapes Constraint Language (SHACL), `https://www.w3.org/TR/shacl/`
29. W3C Community Group: Decentralized Identifiers (DIDs) v0.12, `https://w3c-ccg.github.io/did-spec/`
30. Wood, D.: Ethereum: a Secure Decentralised Generalised Transaction Ledger (2014)
31. Wylot, M., Cudré-Mauroux, P., Hauswirth, M., Groth, P.: Storing, Tracking, and Querying Provenance in Linked Data. IEEE Transactions on Knowledge and Data Engineering **29**(8), 1751–1764 (Aug 2017). https://doi.org/10.1109/TKDE.2017.2690299
32. Wüst, K., Gervais, A.: Do you Need a Blockchain. In: 2018 Crypto Valley Conference on Blockchain Technology (CVCBT). vol. 2017, pp. 45–54 (2018)
33. Yli-Huumo, J., Ko, D., Choi, S., Park, S., Smolander, K.: Where Is Current Research on Blockchain Technology?—A Systematic Review. PLOS ONE **11**(10), e0163477 (Oct 2016). https://doi.org/10.1371/journal.pone.0163477