iLOD: InterPlanetary File System based Linked Open Data Cloud*

Jamal A. Nasir[0000−0002−7866−0020] and John P. McCrae[0000−0002−7227−1331]

Data Science Institute, National University of Ireland Galway, Ireland
{jamal.nasir,john.mccrae}@insight-centre.org

Abstract. The proliferation of the World Wide Web and the Semantic Web applications has led to an increase in distributed services and datasets. This increase has put the infrastructural load in terms of availability, immutability, and security, and these challenges are being failed by the Linked Open Data (*LOD*) cloud due to the brittleness of its decentralisation. In this work, we present *iLOD*: a peer-to-peer decentralized storage infrastructure using the InterPlanetary File System (*IPFS*). *iLOD* is a dataset sharing system that leverages content-based addressing to support a resilient internet, and can speed up the web by getting nearby copies. In this study, we empirically analyze and evaluate the availability limitations of *LOD* and propose a distributed system for storing and accessing linked datasets without requiring any centralized server.

Keywords: Linked Open Data · Data preservation · Blockchain· IPFS

1 Introduction

The World Wide Web and the Semantic Web are managing a network of distributed services and datasets in a decentralized way. An increasing number of datasets are available as Linked Data, also called the Linked Open Data Cloud. But because of this decentralization, sustainability of data is one of the key issues that need to be addressed. While the Linked Open Data Cloud does not have a single point of failure, it also has no mechanism for redundancy so that individual failures gradually decrease the quality of the cloud. Building a sustainable dataset platform is essential for research, and in the past we have seen so many linked data projects (like Laundromat [1]) that went offline once the project's funding was completed. To avoid such failures, blockchain has emerged as a distributed technology to ensure the availability, immutability, and security of data [4]. Currently, there are many highly topical blockchain based applications in cloud computing, and these applications range from information security [14], internet of things (IoT) [11], to the health care domain [7].

Decentralized storage (storage of data independently on multiple nodes of the network in the form of a distributed network) is one of the sustainable solution of datasets storage. A Peer 2 Peer (P2P) architecture, called the Interplanetary File System (*IPFS*),

* Supported by European Union's Horizon 2020 research and innovation programme under grant agreement no. 825182, Prêt-à-LLOD.

has emerged as a solution to store and access applications, files, websites, and data [2]. *IPFS* uses content-based addressing, instead of location-based addressing, which makes the application's physical location transparent and ensures the content remains unique through all nodes.

iLOD capitalizes Header, Dictionary, Triples (*HDT*) format and *IPFS* technology to ensure data preservation by storing datasets securely across multiple locations, and saves in file formats that will likely have the greatest utility in the future [10]. Sharing data in *iLOD* enables data sustainability since *iLOD* facilitates accessing a large number of datasets that were previously limited to dedicated data servers.

The rest of the paper is organized as follows: Section 2 discusses the related work in the field of linked data and outlines scope for further research regarding the limitations of linked data. Section 3 describes the proposed *iLOD* architecture. Finally, the paper concludes by providing a discussion of possible further research.

2 Related Work

A huge increase in distributed services and datasets resulting in a wide range of applications across different sectors [5], but this increase poses challenges in terms of data sustainability and preservation. Many research works introduce efficient data preservation to sustain access to data and files [8,12]. Blockchain has emerged as a distributed technology to ensure data preservation so that data is stored both securely and across multiple locations [6]. *IPFS* is a P2P distributed system to store and access applications, files, websites, and data, using content-based addressing instead of location-based addressing (like domain name, IP address, the path to the file, etc.). The use of content-based addressing enables *IPFS* to uniquely identify the content. Peers (located anywhere in the world) are responsible for providing content as fast as possible. As addressing is content based, all peers would only have the same content copy of the data. This unique addressing makes it easy to link content via directed acyclic graphs (DAGs). This content addressing and linking makes content discovery easy via distributed hash tables (DHTs). As IPFS could be seen as a single BitTorrent swarm, it is employed extensively nowadays by many blockchain projects such as securely sharing personal health records [13], smart transportation systems [15], and reliable image sharing [12].

2.1 Limitations of Linked Data

One of the central challenges associated with linked data is the increasing obsolescence of datasets. Polleres et al.[10] propose the use of *HDT* and *VoID* metadata as a way to improve the quality of the linked data cloud; however, we believe that this is only a partial fix to the challenges facing the linked data cloud. The cloud diagram[1] is one of the main visualizations of the linked data cloud, and is now provided along with extensive metadata. One of these metadata elements is an analysis of the availability of individual datasets as tested by the platform. The results, as summarized in Table 1, report that about three quarters of the datasets are still available in some way; however,

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

	Available	Not Available	Percentage
Datasets	1091	356	75.4%
Links	3344	2239	59.9%
Full Downloads	208	153	57.6%
Other Downloads	2369	1136	67.6%
Examples	499	605	45.2%
SPARQL	268	345	43.7%
application/json	10	14	41.7%
application/ld+json	8	4	66.7%
application/n-triples	8	7	53.3%
application/octet-stream	55	87	38.7%
application/pdf	13	0	100.0%
application/rdf+xml	300	58	83.8%
application/trig	3	1	75.0%
application/xml	40	5	88.9%
application/zip	154	7	95.7%
text/html	1338	227	85.5%
text/plain	44	11	80.0%
text/turtle	263	103	71.9%
text/xml	65	9	87.8%

Table 1. Availability of data in the LOD cloud based on the 27/7/2020 metadata collected at `https://lod-cloud.net`

the results for the links that are still available are less positive, the vast majority of which are categorized as 'other downloads', a category used for indirect links which are often just HTML pages describing the dataset. Direct access to the full data of the datasets is provided by only 25.0% of datasets in the cloud and of those nearly half are no longer available. Similarly, for 'examples' which represent links to a single URL published as linked data (i.e., using content negotiation), we see that most of these resources are unavailable as is also the case for SPARQL endpoints. This shows some of the challenges with publishing data using linked data and maintaining these endpoints for a long time.

Table 1 also presents the availability and total amount of links organized by media type. Most of the major RDF datatypes available in the cloud are Turtle (363 links) followed by RDF/XML (358 links), other formats such as N-Triples (55), JSON-LD (24) and TriG (4) are much less popular, although it is noteworthy that N-Triples are likely much more popular, but as the MIME type was only approved recently and N-Triples are frequently compressed, data in this format may be listed under other media types, such as `text/plain` or `application/octet-stream`. Further, there are currently no links indicating the use of *HDT*. We also see that the most common method of publishing data[2] is ZIP with 95.7%. This fits with the authors' experience in running numerous linked data sites; it is hard to maintain these datasets over the long term, when combined with complex methodologies such as content negotiation

[2] Ignoring formats used for documentation such as HTML and PDF

and SPARQL querying, while a simple dump of the data is the best way to ensure that there is long-term availability of datasets. Likewise, *HDT* takes much less time than traditional techniques to download and start querying a dataset [9].

Up to this point, our analysis has primarily agreed with Polleres [10], however with the reservation that this is a format that will create significant barriers to the usage and publication of data as this format is not widely known or adopted. Furthermore, it is essential to address the challenge of broken links. By the nature of linked data, data is provided with links to other datasets, which are given using HTTP URLs according to the second linked data principle [3]. However, HTTP URLs include a domain, which requires a continuous financial commitment to maintaining the dataset, as the domain must be paid for each year. Eventually, many of these domains will become derelict and those broken links are in the long term a fundamental design flaw of linked data. There are solutions to this such as persistent URLs (e.g., as provided by Purl.org), although a simpler solution would be to enable *mirroring* of datasets so that targets of links are still available, even after the original owner ceases to maintain the datasets. However, this is not possible in linked data due to the fact that in order to update a dataset's URL all incoming links for all datasets that refer to that dataset must also be updated, which is not practical. Therefore, replacement of this brittle location-based addressing with resilient content-based addressing is a new need of the internet. In this work, we use the *IPFS* content-based addressing technique, decentralizing the web itself to give permanence to linked open data.

3 Methodology: *iLOD* Architecture

In this section, we describe the *IPFS* solution for *LOD* datasets. The rationale for this idea is derived from the Table 1. *iLOD* mainly comprises of four parts: Data acquisition, format conversion, *IPFS* addition, and cloud generation (see Figure 1).

Data Acquisition: To allow for maximum flexibility in combining and reusing *LOD*, we consider the *LOD* cloud as a dataset information provider, and the *LOD* Laundromat dump as a data provider. Overlapping datasets are collected. This configuration is for demo only, but any dataset can be used to add in *iLOD*.

Format Conversion: *iLOD* converts all collected datasets into *HDT* format. This compression makes datasets more smaller, while maintaining (and even improving) search and browse operations without prior decompression.

IPFS Addition: The *HDT*-converted datasets are then added to the *IPFS* system to generate a cryptographic hash content identifier (*CID*) for each dataset. *CID* label is a content-based address. Once *CID* has been generated, the dataset is available on the *IPFS* network and ready to share with everyone. Anyone having the *CID* can get content without relying on the (temporary) location of the content. Moreover, this can be used as the target of links by using *IPFS* URLs such as:

```
ipfs://QmPUstQJFCb2yeQNkDjYoq9xL2ftuZ9jpDgQipFKTXfnsE/file.hdt
```

A potential drawback of using these IDs is that content negotiation cannot be implemented. This could easily be fixed by providing good metadata in the *HDT* file to provide alternative versions of the data and backlinks to previous versions that can be

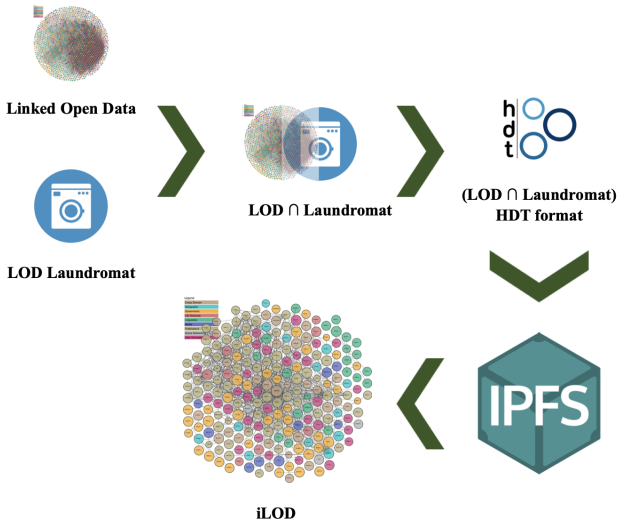

Fig. 1. The architecture of iLOD

easily compiled to provide the most up-to-date version of a dataset. The good thing about *IPFS* is that if the same dataset is added to two different *IPFS* nodes, using the same settings, it will produce exactly the same *CID*, so redundancy issue can also be resolved. A (distributed) index of available datasets can be used to track all datasets and hence find the most recent version.

iLOD Cloud: Generated *CID*s are then added to the *iLOD* cloud. The *iLOD* cloud [3] is a loosely coupled collection of *IPFS* datasets. As the *iLOD* cloud is built on top of the *LOD* cloud, all the advantages and information of *LOD* are also available for *iLOD*. Just like any P2P application, *iLOD* requires active participation in the network in the form of 'pinning' (mirroring) the content, however it is robust to the disappearance of the original data publisher in the network unlike the current *LOD* cloud.

4 Future Work and Conclusions

We have introduced *iLOD* – an *IPFS*-based Linked Open Data that allows users to easily add new datasets or download already added datasets. We suggest extending the linked data principles to require the publishing of linked data as *HDT*, but flexibility can be added if the data is provided with its metadata in a single file following a standard. We also propose a revision of the second principles from "Use HTTP URIs so that people can look up those names" to "use stable, content-based identifiers such as *IPFS*/*CID* URLs to make links so that people can find the data forever." With this, we hope to provide a sustainable platform for *LOD* storage and sharing. In future, we want to extend this work by automatically finding metadata of datasets and their relations with each other in Laundromat dump and then add these datasets to *iLOD* cloud.

[3] `https://lod-cloud.net/#ipfs`

References

1. Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: Lod laundromat: A uniform way of publishing other people's dirty data. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., Goble, C. (eds.) The Semantic Web – ISWC 2014. pp. 213–228. Springer International Publishing, Cham (2014)
2. Benet, J.: IPFS-content addressed, versioned, P2P file system. arXiv preprint arXiv:1407.3561 (2014)
3. Berners-Lee, T.: Linked Data - Design Issues (2006), https://www.w3.org/DesignIssues/LinkedData
4. D'Angelo, G., Ferretti, S., Marzolla, M.: A blockchain-based flight data recorder for cloud accountability. In: Proceedings of the 1st Workshop on Cryptocurrencies and Blockchains for Distributed Systems. pp. 93–98 (2018)
5. Declerck, T., McCrae, J.P., Hartung, M., Gracia, J., Chiarcos, C., Montiel-Ponsoda, E., Cimiano, P., Revenko, A., Sauri, R., Lee, D., et al.: Recent developments for the linguistic linked open data infrastructure. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 5660–5667 (2020)
6. Kim, H.M., Laskowski, M.: Toward an ontology-driven blockchain design for supply-chain provenance. Intelligent Systems in Accounting, Finance and Management **25**(1), 18–27 (2018)
7. Kuo, T.T., Kim, H.E., Ohno-Machado, L.: Blockchain distributed ledger technologies for biomedical and health care applications. Journal of the American Medical Informatics Association **24**(6), 1211–1220 (2017)
8. Li, H., Zhu, L., Shen, M., Gao, F., Tao, X., Liu, S.: Blockchain-Based Data Preservation System for Medical Data. Journal of Medical Systems (2018). https://doi.org/10.1007/s10916-018-0997-3
9. Martínez-Prieto, M.A., Gallego, M.A., Fernández, J.D.: Exchange and consumption of huge rdf data. In: Extended Semantic Web Conference. pp. 437–452. Springer (2012)
10. Polleres, A., Kamdar, M.R., Fernández, J.D., Tudorache, T., Musen, M.A.: A more decentralized vision for linked data **11**(1), 101–113 (Jan 2020), http://semantic-web-journal.net/content/more-decentralized-vision-linked-data-0, sWJ 10-years special issue
11. Wang, X., Zha, X., Ni, W., Liu, R.P., Guo, Y.J., Niu, X., Zheng, K.: Survey on blockchain for internet of things. Computer Communications **136**, 10–29 (2019)
12. Wong, Z.K., Heng, S.H.: Blockchain-Based Image Sharing Application. In: Communications in Computer and Information Science (2020). https://doi.org/10.1007/978-981-15-2693-0_4
13. Wu, X., Han, Y., Zhang, M., Zhu, S.: Secure Personal Health Records Sharing Based on Blockchain and IPFS. In: Communications in Computer and Information Science (2020). https://doi.org/10.1007/978-981-15-3418-8_22
14. Zhang, R., Xue, R., Liu, L.: Security and privacy on blockchain. ACM Computing Surveys (2019). https://doi.org/10.1145/3316481
15. Zichichi, M., Ferretti, S., D'Angelo, G.: A Distributed Ledger Based Infrastructure for Smart Transportation System and Social Good. In: 2020 IEEE 17th Annual Consumer Communications and Networking Conference, CCNC 2020 (2020). https://doi.org/10.1109/CCNC46108.2020.9045640