A Data Market with Decentralized Repositories

Bernd-Peter Ivanschitz[1]
bernd.ivanschitz@researchstudio.at

Thomas J. Lampoltshammer[2]
thomas.lampoltshammer@donau-uni.ac.at

Victor Mireles[3]
victor.mireles-chavez@semantic-web.com

Artem Revenko[3]
artem.revenko@semantic-web.com

Sven Schlarb[4]
sven.schlarb@ait.ac.at

Lőrinc Thurnay[2]
loerinc.thurnay@donau-uni.ac.at

[1]Research Studios Austria, Austria
[2]Danube University Krems, Austria
[3]Semantic Web Company, Austria
[4]AIT Austrian Institute of Technology, Austria

Abstract

In the current era of ever growing data volumes and increased commercialization of data, an interest for data markets is on the rise. When the participants in this markets need access to large amounts of data, as necessary for big data applications, a centralized approach becomes unfeasible. In this paper, we argue for a data market based on decentralized data repositories and outline an implementation approach currently being undertaken by the Data Market Austria project.

1 Introduction

Recently the European Commission has published a study[1] on the potentials of data economy. The study has found that 6 million people in Europe worked in data-related jobs in 2015 and 6.16 million in 2016. As far as medium-term developments are concerned, it is estimated that under a high-growth scenario, the number of data workers in Europe will increase up to 10.43 million, with a compound average growth rate of 14.1% by 2020.

The data industry as a whole comprised approximately of 255,000 data companies in 2016 in the EU. According to the high growth scenario forecast, this figure will increase to 359,050 by 2020 with a compound annual growth rate of 8.9%. The overall value of the data economy grew from the 247 billion in 2013 to almost reaching 300 billion in 2016. According to the estimates of the data market monitoring tool, the value of the data economy in 2016 was worth nearly 2% of the European GDP. By 2020, the EU data economy is expected to increase to 739 billion with an overall impact of 4% on the EU GDP. It is thus fair to say that a healthy data economy can aid in ensuring sustainable employment and growth and thereby societal stability[HL17].

These figured reflect the fact that data has become an important asset in nearly every industry sector[MCB+11]. While the data ecosystem is usually though of as dominated by a few big players, we believe this condition is detrimental both from a competitive standpoint, as well as by the technical limitations that centralizing large amounts of data implies.

Importantly, now that big data technologies are widespread and easily deployable, the volumes of data that gain value in the data economy is also growing. In order to market data that is usually unaccessible because of its sheer volume, it is necessary to come up with decentralized data repositories. This allows for the separation of transactions in the market from actual data transactions. Needless to say, this necessitates that the operation that the buyer of data wishes to perform on the data assets be also executed in a decentralized manner.

In particular, we envision the following features in such a decentralized data market.

[1] https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy

- **Data is decentralized across different repositories** This will counter the existence of so called dark data and the associated dark data lakes [CIK+16]. These phenomena describe the fact that companies or organizations are only able to identify and to utilize a fraction of their data, due to issues related to their inherent business processes, accessibility to data, as well as due to missing knowledge. Furthermore, it will enable the inclusion into the data economy, of data assets that were not originally devised for commercializations and whose continuous use is necessary in other applications.

- **Data processing services are deployed in the infrastructure as needed** In the age of big data and high levels of data heterogeneity, a flexible big data capable infrastructure has become imperative [dSddF+16]. This circumstance is also recognized by the DMA infrastructure, thus offering flexible means of integration regarding data processing services. In case where, e.g., users have a large amount of data residing within their own infrastructure, the DMA provides the possibility to connect external nodes for improved integration (see 1). These external nodes include harvesting services, metadata mappers, as well as blockchain-based data management services to facilitate a high level of connectivity. In consequence, this approach also significantly increases the scalability of the envisioned data market.

- **There is a unified catalogue of data sets and services available on the market** This is necessary for the simple reason that data assets have to be discoverable from a single point of entry. In contrast with distributed (or peer to peer) catalogues, a single catalogue enables the comparison of data assets present in different repositories. This in turn, allows for greater metadata quality by identification of duplicates, increases the power of recommendation systems and allows applications that access several datasets in several infrastructures to be orchestrated. Finally, a single catalogue is easier to connect to other centralized services, in particular to proprietary vocabularies for annotation.

- **There is a distributed, smart contracting system that orchestrates transactions in the market.** The Data Market Austria project decided to employ Ethereum [Woo14] as core component to realize the intended blockchain-based smart contract environment. Within Ethereum, a smart contract is expressed as code fragment that is situated on the blockchain. The associated Ethereum contract is hosted on each individual note within the network. This Ethereum contract comes in form of byte-code, which is executed on the employed Ethereum Virtual Machine (EVM). DMA has opted for the programming language Solidity [Dan17] to formalize the contracts. The underlying program is triggered via the submission of a transaction towards the recipient, paired with the account-wise exchange of Ether according to the actual contract. The Data Market Austria is based on a private Ethereum instance, thus, entities on the platform do not actively use the inherent currency, yet, have to provide the mandatory amount of "gas" to make the transaction/execution possible. Each node within the network will feature the required means in terms of resources to cover necessary operations. These include: i) membership voting for managing participation within DMA; ii) data asset contract: negation processes regarding conditions for accessing and using datasets; iii) service contract: negation processes regarding conditions for accessing and using services.

2 The DMA Implementation

The Data Market Austria (DMA) is a decentralized network of participating (or member) nodes in the sense that there is no central, authoritative location or group that fully controls the market. Nodes are governed by organizations which contribute to the data market by offering their products in form of datasets or services to customers of the DMA. Service providers can at the same time be customers consuming datasets or services of the DMA to offer their own added-value services. This ecosystem will support a full spectrum of data, from open to proprietary data, and will enable innovative business models which can involve any of the potential players.

Each participating node must implement a defined set of services and mandatory standard interfaces. These are, for example instances of a *Data Crawler* a *Metadata Mapper*, a *blockchain peer*, and *Data Management* and *Storage* components. Together with a common conceptual model, these standard interfaces represent the basis to enable interoperability with regard to the use of datasets in the DMA.

The gateway to this decentralized network of nodes containing data and providing services is the *DMA portal* which, while not hosting any data or providing major services, collects information from all nodes to keep an up to date catalogue of available datasets. The node running the portal is denoted as the *Central Node*. A central DMA node is needed to provide a single user-facing window, however, in case the operator of the central node shuts down, it can be rebuilt by another entity, guaranteeing continued operation for the DMA network.

This is where the blockchain technology comes into play. Since blockchains are inherently decentralized: regarded as a a peer-to-peer network of nodes which do not necessarily trust each other, they enable sharing of information between members of the network in a transparent – and with certain limitations also tamper-proof – way. Each member node that is running the blockchain component "knows" about other peers. The DMA members must therefore be able to recreate central services such as the catalogue or user management, using an alternative infrastructure or cloud service provider, should they be shut down or disabled. The information required for recreating these is contained in the immutable and decentralized blockchain.

A Semantic Catalogue for a Data Market with Decentralized Repositories

The trend of collecting data and storing it locally or globally has been going on for the last decade. Rarely has the true value of the data being used or explored. One reason for this is that the data is not made available or accessible for applications or even other datasets to combine it with. Even if companies decide to share or sell there data the structure of the data is often not comparable with other sources which could lead problems. To release the full potential of the data it has to be made easily discoverable and searchable. The use of international data standards, like DCAT, can help with these problems by specifying the descriptive metadata files and the data structure.

To tackle these problems, the DMA use two strategies. First, a global standard is used for the DMA, which is selectively adapted for all the use cases of the DMA. Second, to ensure that also data can be processed that is not in the DMA standard format, interfaces are provided to map the data for the DMA. Especially the second step is essential to ensure an interconnectability with decentralized data repositories, since we can not guarantee that the data is comparable with our standard out of the box.

The DMA metadata catalogue is based on DCAT-AP, the DCAT application profile for data portals in Europe[2] and extends the schema for DMA use cases. This standardization enables future cooperation with international data portals and ensures that the DMA is easily accessible for cooperating companies with a certain data quality standard. The extension focuses on the business use case of the DMA and covers topics like price modeling and dataset exchange, not present in the original DCAT-AP catalogue which was designed for describing public sector datasets. The *priceModel* predicate, for example, allows us to handle the transaction fees for commercial datasets that are being made available in the DMA. The *serviceLevelAgreement* predicate allows to model the condition of a service contract in more details. Without these adaptations, it would not be possible to realize the core services of the DMA.

In the DMA metadata catalogue, every dataset constitutes an RDF[3] resource. There is a set of predicates that link every resource to different literals, which constitute the values of the metadata fields. These values can be of two types: i) literals, as in the case of Author or Description, or ii) elements of a controlled vocabulary, as in the case of Language or License. These controlled vocabularies enable accurate search and filtering. For example, a user searching for datasets in a specific language can do so by selecting from the list of available languages, in which different spellings or abbreviation of one same language are not relevant. Furthermore, they allow an adequate linking of different datasets. If a license the of a dataset is noted as a URI which is provided by the License developers themselves, there is no ambiguity regarding version of the license. The management of controlled vocabularies is achieved through PoolParty Semantic Suite[4].

Due to the decentralized nature of the DMA, metadata is managed on the nodes and thus its normalization into the DMA standard format should also be performed in a distributed way. Not doing so could potentially turn the pre-processing of metadata for the catalogue into a processing bottle neck, and would disable the possibility of recreating the catalogue should the central node leave the DMA. The decentralized normalization has the additional benefit, in line with archival best practices – in particular those adhering to the OAIS model[5] – that datasets and

[2] https://joinup.ec.europa.eu/release/dcat-ap-v11

[3] https://www.w3.org/RDF/

[4] https://www.poolparty.biz/

[5] Reference Model for an Open Archival Information System (OAIS); Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf, version 2 of the OAIS published in June 2012 by CCSDS as "magenta book" (ISO 14721:2012).

related metadata are grouped together [Bru11, p. 129], [Day03, p. 6].

The decentralized metadata normalization requires the separation of the different steps of the metadata processing workflow. As illustrated in Fig. 1 and detailed below, it is assumed that the user has descriptive metadata for each of the corresponding datasets, and that they are familiar with the structure of this metadata. Additionally to enabling the decentralized data market, these steps support two additional use cases: on the one hand, the data provider who has small amounts of data wants to directly upload it to the central node, and, on the other hand, the DMA itself indexing publicly available data.

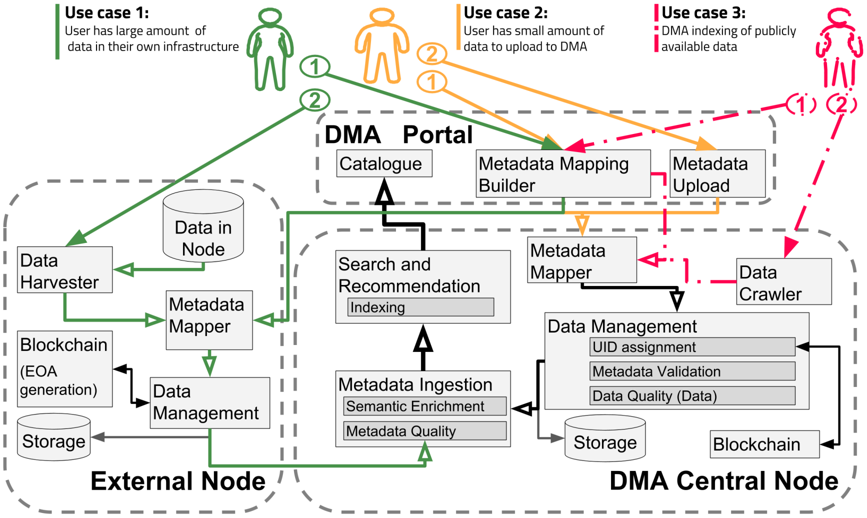

Figure 1: DMA Metadata Ingestion. There are two ways for user to ingest data into the DMA. The first is by uploading data to a web portal, suitable when the amount of data is small. The second is furnishing a machine on the user's premises as a DMA node, and harvesting the data already included in it. Each of these use cases requires the user to interact with two different components of the DMA in a successive manner, shown with the arrows marked with ① and ②. A third way to ingest data into the DMA is for administrators to configure the harvesting of publicly available data portals.

Flow of metadata from a node to the catalogue

When an organization has large amounts of data that it wishes to make available in the DMA, it must not send all of it, nor all of its metadata, to the central DMA infrastructure. Instead, it can instantiate a DMA node in the organization's infrastructure, in which all the processing of data and metadata will take place.

In this workflow, denoted with green arrows in Fig. 1, the node's administrator must first upload a sample of the metadata of their data in JSON or XML into the *Metadata Mapping Builder*. This tool, which is part of the DMA portal, allows a user to configure which of the fields in their metadata file correspond to which fields in the DMA core vocabulary. In a sense, it is a graphical tool to generates XPath or JSONPath expressions. The result is saved in an RDF file that follows the RML specification[DVSC+14]. This file, called a *mapping file*, contains instructions on how to convert any XML (or JSON) file with the same structure into a set of triples.

With the mapping file produced with the Metadata Mapping Builder, the user can return to their own infrastructure and execute the second step. This step consists of inputting the mapping file into the *Data Harvesting Component*, which is part of the basic components of all DMA nodes. This finds, after configuration, the different datasets within the node. The metadata file of each dataset is sent to the *Metadata Mapping Service*, which uses the mapping file created in the first step to generate, for each dataset, a set of RDF triples (serialized in Turtle format). Afterwards, the dataset, its original metadata, and the corresponding RDF are ingested into the *Data Management* component which takes care of the packaging, versioning and assignment of unique identifiers to all datasets, whose hashes are furthermore registered in the Blockchain. All of these steps take place in the user's node.

When the process described above is finished, the node's Data Management component publishes, through a ResourceSync[6] interface, links to metadata files in RDF format of recently added or updated datasets. This way, the node's metadata management is decoupled from the process of incorporating metadata into the DMA catalogue.

In the DMA's central node, the *Metadata Ingestion* component constantly polls the ResourceSync interfaces of all registered nodes, and when new datasets are reported, harvests their RDF metadata which, let us recall, already complies with the DMA metadata vocabulary. This metadata is then enriched semantically. The enrichment is based on EuroVoc[7], which is used in DMA as the main thesaurus. EuroVoc contains 7159 concepts with labels in 26 languages.

For adding the enrichment to the metadata, stand-off annotations are used, i.e. the URIs of the extracted concepts are stored separately and the original titles, description and tags are not modified. These annotations are done using the NLP interchange format [HLAB13]. The predicate "nif:annotation" is used to provide a reference to the knowledge base.

The mapped and enriched metadata is then ingested into the *Search and Recommendation Services*. The high quality of the metadata and its compliance to the chosen scheme guarantees that the datasets and service are discoverable by the users of DMA. Moreover, the usage of the unified vocabularies to describe various attributes of the assets enable more convenient and sophisticated search scenarios such as faceted search or ordered attribute selection. The semantic enrichment is useful for the recommendation service that can, for example, provide better similarity assessments based on semantic comparison.

It is relevant to note that, while a blockchain is available in the DMA as a shared ledger, the possibility of using it also to store metadata of datasets in a fully replicated manner [BSAS17] was discarded for several reasons. First, it was assumed that metadata is changed frequently – e.g. when correcting typos, creating new versions, assigning offers, etc. – and there is actually no need to have a transparent, tamper-proof record of these kind of changes. Second, there is no need to share information regarding each single metadata edit and propagate these changes to all member nodes across the network. Instead, it was considered to be sufficient to capture selected events, such as the publication of a dataset, which are explicitly shared with other member nodes. Third, even though metadata files are relatively small compared to files contained in datasets, the option to use the private Ethereum platform to store metadata files – including related versions created when metadata is changed – would be inefficient in terms of the use of network and storage resources, as data would need to be completely replicated across the whole network. Fourth, the DMA's Search and Recommendation Services use a triple store to allow accessing and querying metadata in an efficient way. The blockchain would not be an appropriate metadata store in this sense.

3 Conclusions

Initiatives for sharing data have now existed for years in many different science domains, such as genome research, geology, or astronomy, just to name a few. Supporting such initiatives with the vision of semantic web standards, in principle, provides the means to create a decentralized, collaborative, interlinked and interoperable web of data [AH12]. In this paper, we have outlined the relevance of metadata for doing the first necessary step to enable a shared data market for Austria: access to multiple distributed repositories through a central portal providing a reliable and consistent basis in terms of normalized and semantically enriched metadata. This serves as the basis for efficient search and recommendation functionalities backed by a central catalogue. However, this only builds the necessary basis. The next step bears the potential to unleash the real power of the market by enabling the *use of aggregated data across distributed repositories*. For this, descriptive metadata, as required for cataloguing, is not sufficient. It is necessary to specify datasets in a way that connectors in data processing software can be instantiated using these descriptions, while simultaneously allowing for effective and transparent contracting mechanisms.

Acknowledgements

The Data Market Austria project is funded by the "ICT of the Future" program of the Austrian Research Promotion Agency (FFG) and the Federal Ministry of Transport, Innovation and Technology (BMVIT) under grant no. 855404

[6] http://www.openarchives.org/rs/1.1/resourcesync

[7] http://eurovoc.europa.eu/

References

[AH12] Sren Auer and Sebastian Hellmann. The web of data: Decentralized, collaborative, interlinked and interoperable. In *LREC 2012*, 2012.

[Bru11] Jorg Brunsmann. Product lifecycle metadata harmonization with the future in oais archives. *International Conference on Dublin Core and Metadata Applications*, 0:126–136, 2011.

[BSAS17] Elena Barriocanal, Salvador Snchez-Alonso, and M Sicilia. Deploying metadata on blockchain technologies. pages 38–49, 11 2017.

[CIK+16] Michael Cafarella, Ihab F Ilyas, Marcel Kornacker, Tim Kraska, and Christopher Ré. Dark data: Are we solving the right problems? In *Data Engineering (ICDE), 2016 IEEE 32nd International Conference on*, pages 1444–1445. IEEE, 2016.

[Dan17] Chris Dannen. *Introducing Ethereum and Solidity*. Springer, 2017.

[Day03] Michael Day. Integrating metadata schema registries with digital preservation systems to support interoperability: a proposal. *International Conference on Dublin Core and Metadata Applications*, 0(0):3–10, 2003.

[dSddF+16] Veith Alexandre da Silva, Julio C.S. Anjos dos, Edison Pignaton de Freitas, Thomas J. Lampoltshammer, and Claudio F.Geyer. Strategies for big data analytics through lambda architectures in volatile environments. *IFAC-PapersOnLine*, 49(30):114 – 119, 2016. 4th IFAC Symposium on Telematics Applications TA 2016.

[DVSC+14] Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. Rml: A generic language for integrated rdf mappings of heterogeneous data. In *LDOW*, 2014.

[HL17] J. Höchtl and Thomas J. Lampoltshammer. Social Implications of a Data Market. In *CeDEM17 - Conference for E-Democracy and Open Government*, pages 171–175. Edition Donau-Universitt Krems, 2017.

[HLAB13] Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. Integrating nlp using linked data. In *International semantic web conference*, pages 98–113. Springer, 2013.

[MCB+11] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H Byers. Big data: The next frontier for innovation, competition, and productivity. 2011.

[Woo14] Gavin Wood. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum project yellow paper*, 151:1–32, 2014.