FAIR Linked Data - Towards a Linked Data Backbone for Users and Machines

Johannes Frey
Sebastian Hellmann
lastname@informatik.uni-leipzig.de
Knowledge Integration and Linked Data Technologies (KILT/AKSW) & DBpedia Association
InfAI, Leipzig University
Leipzig, Germany

ABSTRACT

Although many FAIR principles could be fulfilled by 5-star Linked Open Data, the successful realization of FAIR poses a multitude of challenges. FAIR publishing and retrieval of Linked Data is still rather a FAIRytale than reality, for users and machines. In this paper, we give an overview on four major approaches that tackle individual challenges of FAIR data and present our vision of a FAIR Linked Data backbone. We propose 1) DBpedia Databus - a flexible, heavily automatable dataset management and publishing platform based on DataID metadata; that is extended by 2) the novel Databus Mods architecture which allows for flexible, unified, community-specific metadata extensions and (search/annotation) overlay systems; 3) DBpedia Archivo an archiving solution for unified handling and improvement of FAIRness for ontologies on publisher and consumer side; as well as 4) the DBpedia Global ID management and lookup services to cluster and discover equivalent entities and properties

CCS CONCEPTS

• **Information systems** → **Web searching and information discovery**; *Service buses*; **Semantic web description languages**; • **Applied computing** → *Annotation.*

KEYWORDS

FAIR, Linked Data, Dataset discovery, Data management, Data assessment, Ontologies

ACM Reference Format:

Johannes Frey and Sebastian Hellmann. 2021. FAIR Linked Data - Towards a Linked Data Backbone for Users and Machines. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3442442.3451364

1 INTRODUCTION

The FAIR guiding principles [9] were published in 2016 to ideologically drive the adoption of several guiding principles for publishing scientific and open data. While the principles became widely accepted and popular, they lack technical details and clear, measurable criteria.

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3451364

Three years ago a dedicated European Commission Expert Group released an action plan for "turning FAIR into reality" [1] emphasizing the need for defining "FAIR for implementation".

Linked Data as a technology in combination with the Linked Open Data movement and ontologies can be seen as one of the most promising approaches for descriptive and summarizing metadata but also for representing and connecting research results (data) in form of RDF knowledge graphs across the Web. Although many FAIR principles could be fulfilled by 5-star Linked Open Data [5], the successful realization of FAIR principles for the Web of Data in its current state is a FAIRytale.

In this paper, we give an overview on the implementation of four major approaches that tackle individual challenges of FAIR Linked Data and present our technical vision how they can be extended and interact together to get one step closer towards a backbone for FAIR Linked Data.

2 VISION

Our vision in a nutshell is direct and strict: we envision an information space of metadata that is openly licensed[1], openly accessible in its full extent, fully machine-actionable in terms of access and licenses of the data, globally interoperable, but decentral in nature, coping well with changes on the Web and last-but-not-least has automatically verifiable integrity checks to guarantee all of the aforementioned criteria. Our vision is crucial, because it addresses vagueness and "grey areas" and also defends against "white lies" and "label fraud", i.e. systems claiming to be "open" and "FAIR", but have serious limitations hidden in the fine print or the technical implementation. We particular see the danger in these systems as they exploit the good intentions of FAIR without delivering and without any checks or repercussions.

3 FAIR LINKED DATA CHALLENGES

The FAIR guideline [9] lists 15 principles along the dimensions F (findability), A (accessibility), I (interoperability) and R (re-usability). As a first high-level point of criticism, the guideline leaves many open questions w.r.t. **the realization – Challenge C1** and is barely actionable for data publishers. Secondly, the principles are not **automatically verifiable – Challenge C2**, the overall compliance is not measurable and is therefore hardly awardable and claimable without the ability to check for proof or fulfillment. During our

[1] We argue that Metadata should be CC0, e.g. statements such as "Researcher A provides a CSV with full-genetic material of 10000 subjects, retrievable under URL B with license C, access limitation D and privacy-policy E" must be freely exchangeable

endeavor to renew and advance the creation and release of DBpedia knowledge graphs in the last 3 years [3, 4, 6, 7], we identified several challenges of publishing Linked Data in a FAIR way. We denote the challenges in relation to the respective FAIR principle, followed by the letter *C* for quick reference in the remainder of the paper; (meta)data is abbreviated as $\mathbb{M}/\mathbb{D}$:

F - Findability

F1 $\mathbb{M}/\mathbb{D}$ *are assigned a globally unique and persistent identifier.*

F1C Linked Data as a technology uses globally unique, HTTP-accessible identifier schemes and is therefore in its core well-suited to fulfill F1. These schemes, however, are independently managed and heterogeneous and thus need to cope with identifier deduplication, change adaption (e.g. link rot) and cost of persistence.

F2 *Data are described with rich metadata (defined by R1 below).*

F2C Deciding on the richness and granularity of metadata involves multiple trade-offs between being homogeneous but still flexible enough to accommodate a variety of use cases; between being simple to understand and lightweight (especially in terms of scalability for F4) but being verbose and accurate. Finding a consent for a publishing platform that scales and serves data from various domains seems tough to achieve in the long tail. Moreover, granularity can be dependent on the search context / information demand. There is a need to be able to find/query via *multiple granularity layers* of metadata.

F3 *Metadata clearly and explicitly include the identifier of the data they describe.*

F4 $\mathbb{M}/\mathbb{D}$ *are registered or indexed in a searchable resource.*

F4C We argue that "searchable" is not enough in terms of A,I,R. The resource index must be FAIR itself to prevent $\mathbb{M}/\mathbb{D}$ silos. Furthermore, indexing $\mathbb{M}/\mathbb{D}$ relevant for all communities in one place will be hard or impossible to scale. If $\mathbb{M}/\mathbb{D}$ is split across multiple "FAIR data repositories/indexes" that are non-FAIR itself, questions like a) "How to query multiple metadata (catalogs or repositories) in one attempt/unified way to find relevant information?", and b) "How to *access*, integrate and reuse these?" arise.

A - Accessibility

A1 $\mathbb{M}/\mathbb{D}$ *are retrievable by their identifier using a standardised communications protocol.*

A1C Linked Data standards need to be enforced with validation. Even with PURL(s) link rot occurs. The community is dependent on the registrar to take action (change redirection target).

A1.1 *The protocol is open, free, and universally implementable.*

A1.1C Invalid deployments of Linked Data (invalid serialization, incorrect content-negotiation) limit or prevent access.

A1.2 *The protocol allows for an authentication and authorisation procedure, where necessary.*

A2 *Metadata are accessible, even when the data are no longer available.*

A2C Persistence of ontologies is fundamental to Linked Data. Having persistent access to the metadata only is not sufficient.

I - Interoperability

I1 $\mathbb{M}/\mathbb{D}$ *use a formal, accessible, shared, and broadly applicable language for knowledge representation.*

I2 $\mathbb{M}/\mathbb{D}$ *use vocabularies that follow FAIR principles.*

I2C Recursive definition: If FAIR fails (e.g. link rot) for the used vocabulary, then FAIR will fail for the dataset, eventually. Good RDF datasets are typically built on the shoulder of giants (multiple vocabularies, that itself build again on vocabularies) that are not FAIR (yet, cf. [4]).

I3 $\mathbb{M}/\mathbb{D}$ *include qualified references to other* $\mathbb{M}/\mathbb{D}$.

I3C Biunique, persistent entity/concept identifiers are needed to connect and relate knowledge in an efficient and reliable way.

IC Vocabulary mappings, entity interlinking, and value normalization are needed to allow actual (semantic) interoperability.

R - Reusability

R1 $\mathbb{M}/\mathbb{D}$ *are richly described with a plurality of accurate and relevant attributes.*

R1C Defining "accurate and relevant" $\mathbb{M}/\mathbb{D}$ can be highly community specific and lead to many heterogeneous specialized FAIR repositories that are incompatible to each other. A variety of "rich" metadata vocabularies to attract all kinds of communities and prevent isolated solutions needs to be supported.

R1.1 $\mathbb{M}/\mathbb{D}$ *are released with a clear and accessible data usage license.*

R1.2 $\mathbb{M}/\mathbb{D}$ *are associated with detailed provenance.*

R1.2C For reliable and meaningful provenance the used input/referenced data also needs to be FAIR and the implementation needs to be compatible.

R1.3 $\mathbb{M}/\mathbb{D}$ *meet domain-relevant community standards.*

R1.3C General purpose data can be used interdisciplinary and cross-domain. A mechanism to enrich or associate multiple metadata descriptions independent of the registrar (and without registering duplicates in the same or across repositories) needs to be supported.

4 APPROACHES

We present four major approaches that address individual challenges of FAIR data publishing, try to foster adherence to FAIR or aim at actively improving FAIRness. The chapter is summarized in Table 1.

4.1 Databus, DataID & Collections

The development of the DBpedia Databus[2] was driven by the need for a flexible, heavily automatable dataset management and publishing platform for a new and more agile DBpedia release cycle [7]. It is inspired by the idea of transferring paradigms and techniques from software (release) management/deployment to data management, and the Maven Central Repository.

The Databus uses the Apache Maven concept hierarchy *group*, *artifact*, *version* and ports them to a Linked Data based platform, in order to manage data pipelines and enable automated publishing and consumption of data. *Artifacts* form the abstract identity of a dataset with a stable dataset ID and can be used as entry point to discover all *versions*. A *version* usually contains the same set of *files* for each release. These concepts are embedded in the personal IRI space that is issued by the Databus for every user. The full IRI https://databus.dbpedia.org/<publisher>/<group>/<artifact> /<version>/<file> can be used as a persistent ID (F1) for a particular dataset *file* in a particular *version*. *Groups* provide a coarse modularization or bundling of datasets forming (useful) units. The structuring of *files* is flexible (R1.3), an entire knowledge graph can be represented on *artifact* level, but it also could be represented on the level of a *group*, if artifacts are used to store sub knowledge

[2] https://databus.dbpedia.org

graphs of different entity types or domains (publications and author KG, locations and buildings). Additionally, every *file* can have key-value records which allow an other level of granularity as well as addressing and querying for particular files (F2C). Besides, it is useful to partition an *artifact* (R1.3) systematically into a very large set of files e.g. based on the properties (like e.g. date of birth or place of birth).

An extension of the DataID metadata vocabulary [2] for *artifact*, *version*, and *files* (F2, F3) allows for flexible, fine-grained, as well as unified metadata access using SPARQL (A*, F4). Based on `dcat:downloadURL` links in this metadata, Databus file IDs form a stable (but redirectable) abstraction layer (F1) independent of file hosting (similar to w3id.org). The Databus Maven plugin[3] enables systematic registering of datasets with appropriate metadata on the bus. Provenance can be added by specifying Databus IDs of the input data on *artifact* or *file* level (R1.2(C)). The platform provides several features to tackle and improve F+A for data consumers and producers having in mind machines as well as human users. Free-text metadata (i.e. label, short description, markdown documentation) is indexed in Apache Lucene and provided via a search GUI and API (F4).

Moreover, users can create automatically updating or stable catalogs of data assets via so-called Databus collections[4], which encode the information need / asset selection via SPARQL queries. Collections can be created via a faceted browsing UI and used as easy way to track provenance (R1.2). The collection can be reused as dependency specification for a data workflow/pipeline (e.g. the input datasets of a fused Scientific Knowledge Graph) using the Databus client[5].

4.2 Databus Mods & Overlay Systems

Databus metadata is limited to very basic technical information and free-text documentation. We argue that this metadata should remain very lightweight to scale over a huge amount of assets. To be able to associate and find more and richer metadata (F2) in a unified way, we are proposing the novel Databus Mods (from modifications) architecture[6]. Mods are activities, analyzing and assessing the files or the DataID metadata from the Databus, that can provide useful statistics, enrichment or annotations (e.g. online availability/uptime of the file download location, VoID summaries of datasets, semantic concept annotations for research papers, etc.) Mods allow any user or community (R1.3C) to customize and extend the Databus in a virtual way with their own metadata (R1,R1.3) and add consistent metadata layers (R1C) over all or a selection of Databus files. The mod results (typically metadata extensions F2C) are associated via the PROV ontology[7] using the persistent identifiers of the Databus (F3). Instead of registering files with metadata it is possible to register metadata for Databus files (independent of the registrar).

[3]https://github.com/dbpedia/databus-maven-plugin
[4]http://web.archive.org/web/20200920130715/https://wiki.dbpedia.org/blog/new-prototype-databus-collection-feature
[5]https://github.com/dbpedia/databus-client
[6]https://github.com/dbpedia/databus-mods
[7]https://www.w3.org/TR/prov-o/

Every Mod provider can use the reference implementation (R1C) or develop a custom version of the architecture on their own infrastructure. The architecture uses a master-worker approach. The master monitors the Databus for files of interest and distributes them for analysis to workers via a simple HTTP API. The workers produce a mod result for a file and return it to the master. Every Mod master exposes a SPARQL endpoint (F4, A1*) that stores the PROV-O information (that links the Databus file ID with the mod result) and optionally can also index the mod results itself in case it is represented as RDF. As a consequence, (federated) SPARQL queries (F4C) can be used to obtain information about (1a) the existence of additional community related metadata for Databus IDs or (1b) a selection of files based on Databus core metadata or (2) a selection combining both (or even multiple) metadata vocabularies (F2+F4). Moreover, a ReST API in the master supports on-demand requests for the creation of metadata for a given Databus ID or the retrieval of a mod result.

Complex overlay systems can be built on top of the Mod architecture (e.g. a topic-oriented dataset/file search UI based on a combination of classes and properties from VoID schema summaries[8]). Due to the unified way to associate additional metadata, several types of Mods from different communities can be integrated into one very specialized application (F2C) with a fine-grained selection of files and metadata that are needed as input.

4.3 Archivo

DBpedia Archivo[9] [4] is a dedicated publishing agent[10] (built on top of the Databus) and unified solution for handling and improving FAIR for ontologies on publisher and consumer side – in particular with respect to discovery, versioning, access, and quality-control. Archivo uses the non-information URI of the ontology as the basis for Databus IDs (F1). The host information of the ontology's URI serves as the *group* and the path serves as the name for the *artifact*.

We devised four generic approaches to discover OWL and SKOS ontologies to be archived in Archivo; first and most importantly, by vocabulary usage analysis of all RDF assets on the Databus via VoID Mods (I2C); Second, by querying already existing ontology repositories like Linked Open Vocabularies [8]. Moreover, we discover (transitive) dependencies/imports in ontologies from previous iterations of Archivo crawls (I2C). Finally, users can issue automated inclusion requests for missing ontologies via a Web interface.

Subsequent to the aforementioned discovery steps, the best effort crawling tries to download and parse multiple representations of the ontology and tries to work around typical errors occurring w.r.t. incorrect Linked Data deployments (A1.1C). Archivo crawls for new versions of an ontology every day. The crawled ontologies and metadata are persisted on the DBpedia download server[11] (A2C). Creating a mirrored archive of ontology versions such as Archivo is, of course, not infallible. We consider it, however, a sufficiently reliable fall-back to improve persistence of ontologies.

Appropriate metadata usage (R1) in ontologies itself is validated via SHACLtest suites and can encode general validation rules e.g. to

[8]https://www.w3.org/TR/void/
[9]https://archivo.dbpedia.org/
[10]https://databus.dbpedia.org/ontologies/
[11]13 years in existence, backed up by University of Mannheim

check for a referenced well-known open license (R1.1) or basic metadata coverage (every class should have a label) or sub-community-specific requirements (R1.3). Technical correctness of the Linked Data protocol as well as serialization and logical consistency of the ontology are also verified (I1). These tests are aggregated in a 4-star rating to measure the (re)usability of an ontology and to foster adherence to A1+I+R.

Basic metadata contained in the ontology itself (e.g. Dublin Core terms like `dct:title` or `dct:description`) are fed into the Databus Lucene index for better findability (F2-4).

4.4 Global ID management

The Web of Data uses a decentralized approach with owl:sameAs relations to interlink different RDF resources/entities that represent the same thing. However, a lot of effort is required to obtain a global view in order to discover all sources that provide information about a given entity (thing). We developed the DBpedia Global ID Management [3] to create a central link(ing) hub. In a nutshell, it materializes the global view of links formed by several linksets and datasets available on the Web of Data, computes sameAs clusters by deriving connected components, and selects a DBpedia Global ID as a representative for every cluster, which can be used as uniform identifier for all of its equivalent identifiers (F1C). The ID Management works independent of any link discovery tool. Linking results from any approach can be loaded if they are represented as owl:sameAs links. The Global ID lookup service[12] allows to retrieve all known "equivalent" (F4) IDs or IRI references for a global ID or an ID of an indexed dataset (currently DBpedia, Wikidata, Geonames, Musicbrainz, library data). We also provide an analogous Global Property lookup servicewhich is based on `owl:equivalentProperty` links/mappings (IC). As a consequence, a dataset does not necessarily need to provide a variety of qualified references (I3C) to related sources for one entity (which are likely to break if the target identifiers are not persistent), instead references to a (few) source(s) with persistent IDs (e.g. DBpedia or libraries) are sufficient.

5 REQUIRED EXTENSIONS

Given the current state, communities can create ontologies to annotate files and create search overlay systems for particular domains leveraging Databus Mods. Archivo can enhance the FAIRness of these (meta)data ontologies and improve adherence following clear, verifiable rules. Using the ID and Mapping management, semantic interoperability between datasets as well as the metadata vocabularies can be increased.

Following our vision, we propose **FAIRMeS** (cf. C2, from measurement system) and **AutoFAIR** (tooling for C1) as required extensions towards a successful backbone. While Archivo already encodes and validates a set of FAIR criteria for ontologies, FAIRMeS has the purpose to give a holistic view with detailed statistics and measurements about the FAIRness of a Databus asset: As a next step a combination of Archivo (ratings) and the VoID mods could verify and rate its FAIR vocabulary reuse (I2) for an RDF file, whereas the ID management could compute interlinking scores (I3). Enabled by FAIRMeS, future concrete actions (C1) how to improve bad scores

[12]https://global.dbpedia.org/

Table 1: Overview addressed FAIR principles and challenges

	Databus	Mods + Overlays	Archivo	Global IDs
F1	x		x	
F1C				x
F2	x	x	x	
F2C	x	x		
F3	x	x	x	
F4	x	x	x	x
F4C		x		
A1	x	x		
A1C				
A1.1	x	x		
A1.1C			x	
A1.2	x	x		
A2	x			
A2C			x	
I1			x	
I2				
I2C			x	
I3				
I3C				x
IC				x
R1		x	x	
R1C		x		
R1.1	x		x	
R1.2	x			
R1.2C	x			
R1.3	x	x	x	
R1.3C		x		

could be suggested (e.g. add links/mappings to the following relevant datasets/ontologies). AutoFAIR shall recommend and allow to invoke self-deploying dockerized tools, Web services or mod providers that are automatically configured to improve unFAIR parts of a specific Databus asset based on FAIRMeS scores and suggestions (e.g. permanent hosting, entity linking with related datasets, or even semantic concept annotation for CSV files). We envision an interface and marketplace with matchmaking that in fact would increase the (re)use of tools, services, and data.

We would further like to extend the Databus to support upstreaming of dataset mirrors or later also patches from users if datasets are discontinued or gone (A1C), and in general a more FAIR way to announce hosted services (Lucene index, SPARQL endpoint, overlays) for Databus artifacts (F4Cb).

Open questions exist in terms of scalability of the ID management. As mentioned in F4C a global clustering and indexing for all things in the world is hard to imagine. We think that centralized (sub-community) hubs, which support and coordinate linking in particular domains (e.g. biological) or scopes (e.g. entities like buildings, streets) and provide persistent ids, are needed to achieve good levels of correctness and completeness while keeping the complexity for linking on a manageable scale. A data DNS - a hierarchical delegation for search, registration, and link management could help to distribute the workload in a unified and global way. This concept could be extended to all presented systems, forming sub-databuses, sub-archivos, and sub-management systems for mods.

ACKNOWLEDGMENTS

This work was partially supported by grants from the Federal Ministry for Economic Affairs and Energy of Germany (BMWi) to the projects LOD-GEOSS (03EI1005E) and PLASS (01MD19003D).

REFERENCES

[1] Sandra Collins, Françoise Genova, Natalie Harrower, Simon Hodson, Sarah Jones, Leif Laaksonen, Daniel Mietchen, Rūta Petrauskaitė, and Peter Wittenburg. 2018. Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data. https://doi.org/10.2777/54599

[2] Markus Freudenberg, Martin Brümmer, Jessika Rücknagel, Robert Ulrich, Thomas Eckart, Dimitris Kontokostas, and Sebastian Hellmann. 2016. The Metadata Ecosystem of DataID. In *Metadata and Semantics Research*, Emmanouel Garoufallou, Imma Subirats Coll, Armando Stellato, and Jane Greenberg (Eds.). Springer International Publishing, Cham, 317–332.

[3] Johannes Frey, Marvin Hofer, Daniel Obraczka, Jens Lehmann, and Sebastian Hellmann. 2019. DBpedia FlexiFusion the Best of Wikipedia > Wikidata > Your Data. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 11779)*, Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon (Eds.). Springer, 96–112. https://doi.org/10.1007/978-3-030-30796-7_7

[4] Johannes Frey, Denis Streitmatter, Fabian Götz, Sebastian Hellmann, and Natanael Arndt. 2020. DBpedia Archivo: A Web-Scale Interface for Ontology Archiving Under Consumer-Oriented Aspects. In *Semantic Systems. In the Era of Knowledge Graphs - 16th International Conference on Semantic Systems, SEMANTiCS 2020, Amsterdam, The Netherlands, September 7-10, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12378)*, Eva Blomqvist, Paul Groth, Victor de Boer, Tassilo Pellegrini, Mehwish Alam, Tobias Käfer, Peter Kieseberg, Sabrina Kirrane, Albert Meroño-Peñuela, and Harshvardhan J. Pandit (Eds.). Springer, 19–35. https://doi.org/10.1007/978-3-030-59833-4_2

[5] Ali Hasnain and Dietrich Rebholz-Schuhmann. 2018. Assessing FAIR Data Principles Against the 5-Star Open Data Principles. In *The Semantic Web: ESWC 2018 Satellite Events*, Aldo Gangemi, Anna Lisa Gentile, Andrea Giovanni Nuzzolese, Sebastian Rudolph, Maria Maleshkova, Heiko Paulheim, Jeff Z Pan, and Mehwish Alam (Eds.). Springer International Publishing, Cham, 469–477.

[6] Sebastian Hellmann, Johannes Frey, Marvin Hofer, Milan Dojchinovski, Krzystof Węcel, and Wlodzimierz Lewoniewski. 2020. Towards a Systematic Approach to Sync Factual Data across Wikipedia, Wikidata and External Data Sources.. In *Proceedings of the Conference on Digital Curation Technologies.* https://svn.aksw.org/papers/2020/qurator_gfs/public.pdf

[7] Marvin Hofer, Sebastian Hellmann, Milan Dojchinovski, and Johannes Frey. 2020. The New DBpedia Release Cycle: Increasing Agility and Efficiency in Knowledge Extraction Workflows. In *Semantic Systems. In the Era of Knowledge Graphs - 16th International Conference on Semantic Systems, SEMANTiCS 2020, Amsterdam, The Netherlands, September 7-10, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12378)*, Eva Blomqvist, Paul Groth, Victor de Boer, Tassilo Pellegrini, Mehwish Alam, Tobias Käfer, Peter Kieseberg, Sabrina Kirrane, Albert Meroño-Peñuela, and Harshvardhan J. Pandit (Eds.). Springer, 1–18. https://doi.org/10.1007/978-3-030-59833-4_1

[8] Pierre-Yves Vandenbussche, Ghislain Atemezing, María Poveda-Villalón, and Bernard Vatant. 2017. Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web. *Semantic Web* 8, 3 (2017), 437–452. https://doi.org/10.3233/SW-160213

[9] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.