FAIR Data Points Supporting Big Data Interoperability

Luiz Olavo Bonino da Silva Santos[1] — Mark D. Wilkinson [2] — Arnold Kuzniar [3] — Rajaram Kaliyaperumal [4] — Mark Thompson [4] — Michel Dumontier [5] — Kees Burger [4]

1 Dutch Techcentre for Life Sciences and Vrije Universiteit Amsterdam
The Netherlands

luiz.bonino@dtls.nl
2 Center for Plant Biotechnology and Genomics, Universidad Politécnica de Madrid, Spain
markw@illuminae.com

3 Netherlands eScience Center, The Netherlands

a.kuzniar @esciencecenter.nl
4 Biosemantics Group, Leiden University Medical Center, The Netherlands
m.thompson@lumc.nl
r.kaliyaperumal@lumc.nl
5 Stanford Center for Biomedical Informatics Research, Stanford University, USA
michel.dumontier@stanford.edu

ABSTRACT: *With the evolution and widespread adoption of contemporary information technologies, data has taken an increasingly central role in almost all areas of human activity. While on one hand this protagonism of data brings significant benefits for the individuals and organizations, on the other hand it is accompanied by a number of challenges. In the domain of Big Data, Linked Data and Semantic Web, common examples of these challenges include scale, performance, availability, security, diversity, complexity, semantics, manageability and findability. In January 2014, a group of stakeholders involved in research data came together to debate how to further enhance infrastructures to support a data ecosystem that promotes data interoperability and reuse. The main outcome of this meeting was the definition of the so called FAIR Data guiding principles aimed at publishing data in a format that is Findable, Accessible, Interoperable and Reusable by both machines and human users. Since the seminal meeting, the FAIR Data initiative has grown with an increasing number of participant organizations and involved projects. In this chapter we report the progress in the development of one of the components of the FAIR Data infrastructure, the FAIR Data Point (FDP). The FDP is software that, from one side, allows data owners to expose datasets in compliance with the FAIR principles and, from another side, allows data users to discover information about the available datasets and, ultimately, access the underlying data records.*

KEY WORDS: *FAIR Data, data interoperability, Semantic Web, Linked Data, metadata.*

Enterprise Interoperability - Proceedings of the Workshops of the Eighth International Conference I-ESA 2016, pages 1 to n

1. Introduction

The FAIR Data Principles [FAIRDP] propose that scholarly output, including both data and the metadata and workflows that surround them, should be Findable, Accessible, Interoperable and Reusable. The FAIR Principles aim to address the lack of widely-shared, clearly articulated, and broadly applicable best-practices around the publication of the data generated by scientific research. While the history of scholarly publication in journals is long and well-established, the same cannot be said of formal data publication; yet, data could be considered the primary output of scientific research, and its publication and reuse is necessary to ensure validity, reproducibility, and to drive further discoveries. The FAIR Data Principles address these needs by providing a precise and measurable set of qualities a good data publication should exhibit - qualities that ensure that the data is easily discovered, easily evaluated, and maximally reusable.

The principles were formulated after a Lorentz Center workshop in January, 2014, where a diverse group of stakeholders, sharing an interest in scientific data publication and reuse, met to discuss the features required of contemporary scientific data publishing environments. The first-draft FAIR Principles were published on the Force11 website for evaluation and comments by the wider community - a process that lasted almost two years, and resulted in the clear, concise, broadly-supported principles that were recently formally published [Wilkinson *et al*]. The principles support a wide range of new international initiatives, such as the European Open Science Cloud [EOSC] and the NIH Big Data to Knowledge [BD2K], by providing clear guidelines that help ensure all data and associated services in the emergent 'Web of Data' will be Findable, Accessible, Interoperable and Reusable, not only by people, but notably also by machines.

The recognition that computers must be capable of accessing published data autonomously, unaided by their human operators, is core to the FAIR Principles. Computers are now indispensable tools in every research endeavor, as contemporary scientific datasets are large, complex, and globally-distributed, making it almost unfeasible for humans to manually discover, integrate, inspect and interpret them. This (re)usability barrier has, until now, prevented us from maximizing the return-on-investment from the massive global financial support of "big data" research and development projects, especially in the life and health sciences. For this reason, the FAIR Principles put equal weight on both the "FAIR-ness" for humans, and the "FAIR-ness" for machines, through the requirement to use machine-interpretable data formats and controlled vocabularies for (meta)data publications.

While the intent of the FAIR Principles is clear, the implementation details of these principles is (purposely) avoided in their formulation and wording. As such, this important task is left for the data publishers themselves, and the Principles provide no further guidance in the use of technologies.

While the FAIR Data Principles do not suggest or promote any specific technology, this manuscript provides an exemplar implementation of FAIR in the form of what we call a "FAIR Data Point" (FDP). FDPs intend to act as a software layer over data resources, allowing them to expose their content in a strict compliance with the FAIR Principles, for both humans and machines. In this chapter we describe the layered architecture of the FDP solution, focusing on the mechanism by which we provision metadata and data in a discoverable and machine-accessible manner. In addition, we demonstrate using a collection of FDPs exposing a number of datasets to show how they support client applications with findability and accessibility of data. Finally, we inform the next steps in the development of FDPs, their application to existing data repository services such as EUDAT, their integration with other elements of the FAIR Data infrastructure, and conclude with an overview of the accomplishments so far.

This chapter is further structured as follows: section 2 discusses the metadata necessary to provide proper findability to FAIR Data applications in general and FDPs in particular. Section 3 presents the architecture of the FDPs and details the main components. Section 4 presents an example application gathering data from multiple FAIR Data Points and integrating these datasets to satisfy information request from its users. Section 5 discusses the next steps on the development of FDPs and section 6 gives the final remarks and conclusions.

2. Metadata

The scale of the World Wide Web rapidly exceeded the ability of humans to find content of interest by following links, and search engines such as Altavista, Google and Bing indexing the content of the Web to facilitate search became fundamental. The Web of Data intends to provide access to all published datasets; however, discovery of data poses a very different problem compared to the discovery of narrative text. Since individual data elements (for example, numerical cells or columns in a spreadsheet) are highly granular, it is not rational to index such data. Moreover, the nature of the data content cannot be determined by "crawling" the dataset, which prevents its discovery and reuse. Efforts to catalog available data sources are high-level and generally domain-specific such as the Nucleic Acids Research Database Summary [NAR], which catalogs over 1.600 molecular biology databases. Such efforts provide a list of databases within a specific topic, along with additional information such as title, publisher, summary and URL of the database. Although already helpful, these catalogs do not offer true search functionality within the data itself, for example, to determine which databases contain records relating to some entities of interest (e.g. a gene or a disease). Finally, these catalogs are generally the result of manual curation, and cannot be automatically constructed because of the lack of published, machine-searchable metadata.

The first facet of FAIR Data, therefore, addresses issues of data(set) discoverability. Briefly, the Findability facet of FAIR requires that there is sufficient machine-readable and indexable metadata, about every dataset, for a user to decide whether a given data provider has records of-interest. Moreover, beyond finding information about desired data elements, other metadata is required in order to arrive at the final decision to access and reuse the discovered data. These include the perceived trustworthiness of the data creator/publisher, usage restrictions imposed by the license, the representation format of the data, and/or its semantics.

FDPs comply with the FAIR Data Principles by providing metadata that is divided into four complementary layers, namely FDP Metadata, Data Catalog Metadata, Dataset Metadata and Data Record Metadata.

The FDP Metadata layer provides general information about the FDP such as name and description, its provenance and technical details such as a link to the formal description of the required, recommended, and optional metadata elements that software can expect from this FDP Metadata service. The definition of the content of this layer is based on the Open Archives Initiative Protocol for Metadata Harvesting [OAI-PMH] and uses Dublin Core metadata terms [DCT].

The Data Catalog Metadata layer provides information about the list of datasets offered by the FDP. For the representation of the catalog metadata, we adopted the W3C's Data Catalog Vocabulary [DCAT]. In DCAT, a catalog is defined as “a curated collection of metadata about datasets”.

A dataset, as a collection of individual data items, may be available in different formats. For instance, the Human Protein Atlas [HPA] offers access to its datasets through its API and also in XML, RDF and tab-delimited formats. The Dataset Metadata layer provides information about each of the datasets in the FDP's catalog, including the forms in which they are available. As with the Catalog Metadata, we adopted DCAT to specify the metadata elements within the FDP Dataset Metadata. DCAT specifies both the dataset metadata and the metadata about the forms in which the dataset is made available using the concepts of *dataset* and *distribution*, respectively. In DCAT, a catalog is defined as “a collection of data, published or curated by a single agent, and available for access or download in one or more formats” while a distribution is defined as representing “a specific available form of a dataset. Each dataset might be available in different forms and these forms might represent different formats of the dataset or different endpoints”.

The fourth layer of metadata represents the data record, i.e., it provides information about the data items contained in the dataset. This information allows the users to assess what is the actual content of the dataset, by describing the data types represented in the data. According to the FAIR Data Principles, data should be sufficiently well-described and rich that can be automatically linked or integrated as well as utilise shared vocabularies and/or ontologies. These vocabularies are used to provide semantics to the data types they annotate. For instance, a dataset containing genetic data could use the concept of gene from the Systematized Nomenclature of Medicine - Clinical Terms [SNOMEDCT] ontology as a reference for the gene type it presents. In this case, the data record metadata would contain the information that there is a concept named gene, identified by the URI http://purl.bioontology.org/ontology/SNOMEDCT/67271001 and this concept is used to classify an entity present in the dataset.

As expected, the content of different datasets varies and so the data record metadata. Some types of datasets have their content standardised and, since their metadata reflect the content are standardised. An example of such data content standardisation is the Minimum Information About Biobank Data Sharing [MIABIS], which defines what are the data types in a sample collection dataset, their value types and range. Due to the inherent heterogeneity of datasets content, the FDP Data Record metadata layer is described in terms of the standard adopted to describe the given dataset content.

3. Architecture

The FAIR Data Point is software that, on one hand, allows data owners to expose datasets in a FAIR manner and, on the other hand, allows data users to discover properties about offered datasets by means of their metadata and, if license conditions allow, to access the data itself. Although a FDP may be used in any knowledge domain, our focus is on life sciences and, therefore, the examples are concerned with biological datasets.

A key architectural requirement is that FDPs are distributed. It is inconceivable that there will ever be a unique and centralized repository for each scientific data-type. Despite the impracticality of having a huge infrastructure to host all the existing and future data resources, other factors such as legislation, privacy and security risks, performance and transportation capacity requires a distributed environments where a number of large reference data repositories, containing well-curated and integrated core datasets, e.g., the repositories available at the European Bioinformatics Institute (EBI), is combined with smaller distributed data repositories such as different biobanks, datasets or databases created within the scope of research projects.

Currently, the FDP software is developed as a stand-alone Web application. However, it has been designed such that it can be incorporated in other applications by either including the current implementation as a component of a larger application or by extending an existing application to conform with the FDP's specifications such as API, and data and metadata content and formats.

As depicted in Figure 1, FDP has four main components namely the Metadata Provider, the Data Accessor, the Security Enforcer and the Metrics Gatherer. In the following sections, we describe each of these components but focus on the Data Accessor, and the Metadata Provider component, that are the current focus of development. A FAIR Data Point can be accessed by a user through its graphical user interface (GUI) and by computational clients through its application program interface

(API). In our implementation of the FDP, we adopt a RESTful [Fielding] interface for the API.

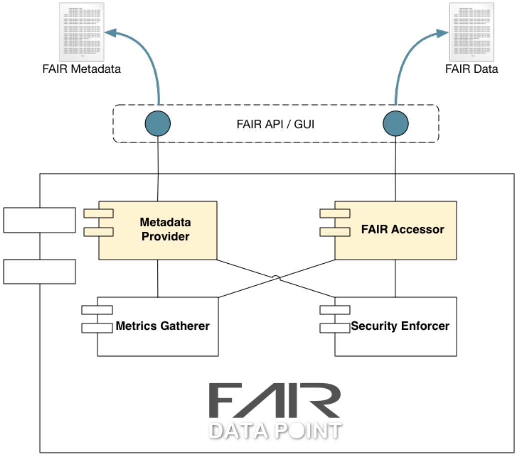


Figure 1 - *FAIR Data Point components*

3.1. *Metadata Provider*

The Metadata Provider is responsible for giving access to the metadata described in Section 2. The component is accessible as a service to the users through its REST API. In our implementation the Metadata Provider service responds to the root URL of the FDP.

The implementation of the services produces a structure of resource paths. However, it is not our goal to specify an arbitrary and monolithic API. To avoid this, we adopt the Hypermedia as the Engine of Application State (HATEOAS) [Fielding] pattern. In short, a HATEOAS API provides information on how to navigate through the API even if the client does not have previous knowledge of the interface. In our case, once the client request information from the root URL, the FDP responds with the FDP Metadata content together with the link to the URL of the next navigational path, the Catalog Metadata resource.

3.2. FAIR Accessor

The FAIR Accessor component provides access to the actual data content of the dataset. In our implementation of the FDP we have adopted the W3C's Linked Data Platform [LDP] standard for the FAIR Accessor. LDP defines a set of rules for HTTP operations on web resources, providing an architecture for read-write Linked Data on the web. LDP is based on HTTP operations such as GET, POST, PUT and DELETE, leveraging from the significant amount of supporting tools and libraries. Another benefit of LDP is its support to deal with large resources by providing the ability to break large datasets into multiple paged responses.

3.3. Metrics Gatherer

The Metrics Gatherer component monitors various aspects the FDP usage. These metrics can be used by the FDP owner to assess the load on its FDP and, thus, giving the owner the required information to adjust the infrastructure accordingly. Moreover, since one of the FAIR Data principles is the possibility for data to be citable, the information about the usage of the published data can derive the equivalent of the impact factor metric used in scientific manuscript publications.

3.4. Security Enforcer

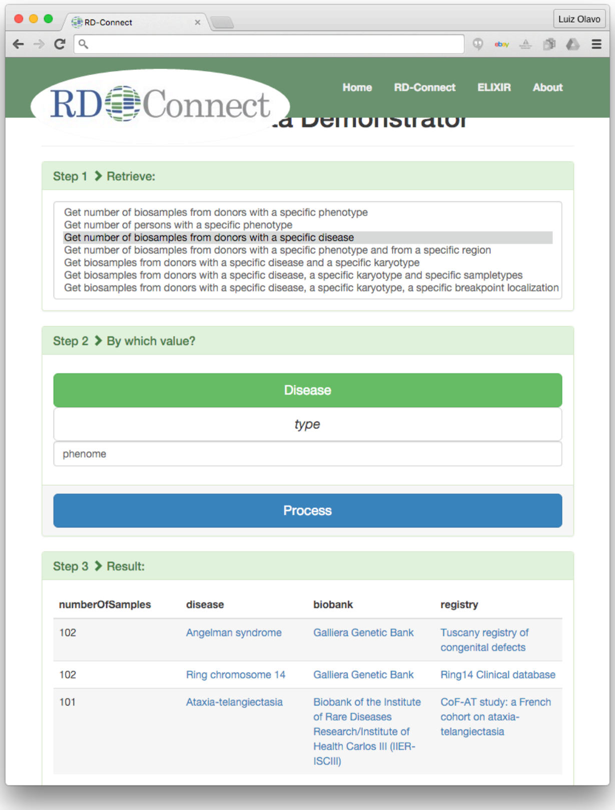
The Security Enforcer component should act as a gatekeeper, protecting the access to the (meta)data from requests that do not comply with the given licenses. It is important to reinforce that the "A" in FAIR Data does not necessarily mean open. It means accessible through specific conditions. And these conditions are defined in the license.

Besides securing the access to the data, the Security Enforcer is also responsible securing the storage and transmission of the data by means of encryption, when necessary. In the case of sensitive data, it is desirable that they are stored in an encrypted form, not being readable even by the owner of the storage facility. For instance, the FDP owner may employ a third-party untrusted cloud provider for the data storage of the data and the provider would not be able to understand the data since they are encrypted.

Another feature for the Security Enforcer is pseudonymization of the data. In the scenario where a FDP contains data about individuals, such as a FDP containing patient records, the data owner may grant access to the data as long as the data do not allow the identification of individuals.

4. Usage example

The FDP is being developed under the collaboration of a number of research projects, namely Open Discovery and Exchange for all (ODEX4all), RD Connect and Biobanking and BioMolecular Resources Research Infrastructure - The Netherlands (BBMRI-NL 2.0) [BBMRI]. In this context the engineering team has created a demonstration application for the interoperability possibilities of FAIR Data using FAIR Data Points to make available data from patient registries and biobanks. The patient registries contain data about patients of rare diseases such as patient identification, the rare disease of the patient, age of onset, examinations, treatments, genetic information and pharmacological therapies. The biobanks collect biological or medical data and tissue samples.



Step 1 › Retrieve:

- Get number of biosamples from donors with a specific phenotype
- Get number of persons with a specific phenotype
- Get number of biosamples from donors with a specific disease
- Get number of biosamples from donors with a specific phenotype and from a specific region
- Get biosamples from donors with a specific disease and a specific karyotype
- Get biosamples from donors with a specific disease, a specific karyotype and specific sampletypes
- Get biosamples from donors with a specific disease, a specific karyotype, a specific breakpoint localization

Step 2 › By which value?

Disease

type

phenome

Process

Step 3 › Result:

numberOfSamples	disease	biobank	registry
102	Angelman syndrome	Galliera Genetic Bank	Tuscany registry of congenital defects
102	Ring chromosome 14	Galliera Genetic Bank	Ring14 Clinical database
101	Ataxia-telangiectasia	Biobank of the Institute of Rare Diseases Research/Institute of Health Carlos III (IIER-ISCIII)	CoF-AT study: a French cohort on ataxia-telangiectasia

Figure 2 - *GUI for the demo FDP client application*

The goal is to demonstrate how by adhering to the FAIR Principles not only data sources from the same type, such as patient registries, can improve their interoperability but also across different types of data sources, such as combining data from patient registries with data from biobanks.

For this demonstration, we have deployed a number of FDPs serving (meta)data from patient registries and biobanks. Then, an aggregation application harvests the metadata from the FDPs, verifies which ones contain data relevant to the application's workflows and retrieves the data from the selected FDPs in a cache. When a user of the application selects one of the available workflows and enters the required parameters, the application queries the cached data and returns the results to the user.

Figure 2 shows the web interface of the aggregation application after the user selected the "Get number of biosamples from donors with a specific disease" workflow and the disease. The results show the number of available biosamples for each given diseases matching the criteria, which biobank stores the samples and from which patient registry the donor has information on. Also in the interface, if the user selects the links represented by the disease names, the semantic annotation of the data is resolved to the landing page of the disease's concept in the associated ontology. For the links represented by the biobank and patient registries names, the user is taken to their respective web pages containing information on how to request the samples or the patient registry data.

5. Current status and next steps

Currently, the FAIR Data Point development is focused on the Metadata Provider and FAIR Accessor. The next step in our development plan is to tackle the security issue by implementing the Security Enforcer, starting, in the coming months, with encryption and pseudonymization technologies that can be used to protect sensitive data such as human health data. After this we will start the development of the Metrics Gatherer component.

Besides the actual software development, in the coming period we will also focus on validating the approach by deploying FDPs in a wider range of application scenarios while monitoring their usage and consequent improvements in data sharing and interoperability. Moreover, we will pursue the engagement of a larger number of stakeholders to discuss the metadata content and to define which of the described metadata elements should be required, suggested and optional.

FDPs have been designed with the intention of facilitating mainly findability and accessibility. In the current phase of our work we are not focusing on the quality of the data being exposed through FDPs. However, in the scope of FAIR Data, work should be carried out to evaluate how the compliance with the FAIR Data Principles impact the data quality and what are the additional approach necessary to guarantee an intended data quality level.

6. Conclusions

The current scenario of data exchange and interoperability and, more specifically regarding scientific data, requires a higher degree of automation. To achieve this improvement in automation in a global scale, one sensible approach is to have the minimal possible set of standards and guidelines. This lightweight approach is the basis of the FAIR Data initiative. The architecture, design and development of the FDP also follows this by defining the use of a small set of standards, namely, OAI-PMH, Dublin Core, DCAT and LDP.

In this chapter we have described the current status of the development of the FDP, its architecture, components and related metadata. We also presented an example of a client application for the FDP developed in the context of a combination of research projects in the area of Life Sciences. Finally, we discussed the next steps for the development of FDP.

References

Big Data to Knowledge [BD2K]. 2015. Web. <https://datascience.nih.gov/bd2k>.

Biobanking and BioMolecular Resources Resarch Infrastructure – The Netherlands [BBMRI]. 2014. Web. <http://www.bbmri.nl>.

Data Catalog Vocabulary [DCAT]. W3C, 2014. Web. <http://www.w3.org/TR/vocab-dcat/>.

Dublin Core metadata terms [DCT]. 2012. Web. <http://dublincore.org/documents/dcmi-terms/>.

European Open Science Cloud [EOSC]. Web. <https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud>.

FAIR Data Principles [FAIRDP], FORCE 11, Sept. 2014. Web. <https://www.force11.org/group/fairgroup/fairprinciples>.

Fielding R. T., “Chapter 5: Representational State Transfer (REST)”, *PhD Thesis*, University of California Irvine, chapter 5, 2000.

Human Protein Atlas [HPA]. Nov. 2015. Web. <http://www.proteinatlas.org/>.

Linked Data Plataform [LDP]. 2015. Web. <https://www.w3.org/TR/ldp/>.

Norlin L., *et al*, “A Minimun Data Set for Sharing Biobank Samples, Information and Data: MIABIS”, *Biopreserv Biobank*, vol. 10 no. 4, august 2012, p. 343-348.

Nucleic Acids Research Database Summary [NAR], Oxford Journals, 2016. Web. <https://www.oxfordjournals.org/our_journals/nar/database/a/>.

Open Archives Initiative Protocol for Metadata Harvesting [OAI-PMH], OAI, 2015. Web. <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

Systematized Nomeclature of Medicine [SNOMEDCT]. 2015. Web. <https://www.nlm.nih.gov/snomed/>.

Wilkinson M., *et al,* “The FAIR Guiding Principles for scientific data management and stewardship”, *Scientific Data*, Nature, doi:10.1038/sdata.2016.18