Assessing FAIR data principles against the 5-star Open data principles

Ali Hasnain[1] and Dietrich Rebholz-Schuhmann[1]

Insight Centre for Data Analytics, National University of Ireland, Galway
firstname.lastname@insight-centre.org

Abstract. Access to biomedical data is increasingly important to enable data driven science in the research community. The Linked Open Data (LOD) principles (by Tim Berner-Lee) have been suggested to judge the quality of data by its accessibility (open data access), by its format and structures, and by its interoperability with other data sources. The objective is to use interoperable data sources across the Web with ease. The FAIR (findable, accessible, interoperable, reusable) data principles have been introduced for similar reasons with a stronger emphasis on achieving reusability. In this manuscript we assess the FAIR principles against the LOD principles to determine, to which degree, the FAIR principles reuse LOD principles, and to which degree they extend the LOD principles. This assessment helps to clarify the relationship between both schemes and gives a better understanding, what extension FAIR represents in comparison to LOD.
We conclude, that LOD gives a clear mandate to the openness of data, whereas FAIR asks for a stated license for access and thus includes the concept of reusability under consideration of the license agreement. Furthermore, FAIR makes strong reference to the contextual information required to improve reuse of the data, e.g., provenance information. According to the LOD principles, such meta-data would be considered interoperable data as well, however, the requirement of extending of data with meta-data does indicate that FAIR is an extension of the LOD (in contrast to the inverse).

Keywords: Linked Open Data (LOD), Linked Open Data 5-star, FAIR data principles

1 Introduction

The advent of the World Wide Web [2] has enabled public publishing and consumption of information on a unique scale in terms of cost, accessibility and size. In the past few years, the linked open data cloud has earned a fair amount of attention and it is becoming the standard for publishing data on the Web [10,5,7].

One of the ambitions behind the linked data effort is the ability to create a Web of interlinked data which can be queried using a unified query language and protocol, regardless of where the data is stored [8]. Core to this achievement

is the adoption of the resource description framework (RDF) as the knowledge representation formalism as well as the SPARQL protocol for the retrieval of data. Ideally, all data is openly available and accessible through a maintained IT infrastructure [11].

Since the idea of linked data is in place, different schemes, proposals and recommendations (with guiding principles) have emerged from researchers and practitioners to make it open and easy to use by others[12]. One such scheme is proposed by Tim Berners-Lee in year 2010 known as "Linked Open Data 5-star [1]". Under the star scheme, one star can be get if the information has been made public under an open licence. More stars denote the quality of data leading to better structure and interoperability for reuse.

While LOD has had some uptake across the web, the number of datasets using this protocol compared to the other technologies is not up to the mark and significantly modest [8]. But whether or not one uses LOD, one does need to ensure that the datasets are designed specifically for the web and for reuse by humans and machines. To provide guidance for creating such data content independent of the technology used, recently the FAIR principles were issued through the Future of Research Communications and e-Scholarship (FORCE11)[3]. The FAIR principles provided by Wilkinson et al., [16] put forth characteristics that contemporary data resources, tools, vocabularies and infrastructures should exhibit to assist discovery and reuse by third-parties through the web. FAIR stands for: Findable, Accessible, Interoperable and Re-usable.

In this paper we want to highlight the overlapping aspects proposed by Linked Open Data 5-star and FAIR principles and present our position on how these two are related .

2 Related Work

Cox et. al, [4] presented a rating system known as the 5* OzNome Data tool[1] which allows users to carry out a self-assessment by classifying the 14 facets into data quality guidelines. These data quality guidelines are provided and derived by FAIR principles (Findable, Accessible, Interoperable, Reusable) along with an additional dimension called Trustable or being Trusted. In doing the self-assessment, one can explore ways in which they are able to improve their data collection and how it is accessed by others. This gives data providers targets to improve their data collection and publishing process.

Related work can also be taken into account from the perspective of quality analysis of datasets. Although this paper doesn't provide insights or comparative quality analysis of existing approaches, but both the FAIR and 5-star principles targets the quality aspects of data.

Yamaguchi et al, presented YummyData [17] which is designed to improve the findability and reusability of Life Science Linked Open Data (LS-LOD) by monitoring the states of their Linked Data implementations and content.

[1] `(http://oznome.csiro.au/5star/)la:20-03-2018`

SPARQLES[2] [14] monitors SPARQL endpoints registered in DataHub to determine their availability, performance, interoperability, and discoverability.

Hasnain et al, [8,9] presented SPORTAL (SPARQL portal) that provides the analysis in terms of discoverability, findability and accessibility aspects of Linked Open Data publish as public SPARQL endpoints.

3 LOD 5-star

Tim Berners-Lee proposed 5-star scheme for Linked Open Data that provide guidelines for data providers and publishers in order to make data more accessible, available and reusable over the Web.

$\star$ make your stuff available on the Web (whatever format) under an open license
$\star\star$ make it available as structured data (e.g., Excel instead of image scan of a table)
$\star\star\star$ make it available in a non-proprietary open format (e.g., CSV instead of Excel)
$\star\star\star\star$ use URIs to denote things, so that people can point at your stuff
$\star\star\star\star\star$ link your data to other data to provide context

In the following subsection we look into the importance of achieving these stars from user and publisher perspective

- ***One*** $\star$: Achieving *one star* means that the user can be able to 1) access the data, 2) consume the data, 3) store it locally, 4) manipulate the data and 5 share the data. Whereas being publisher 1) it is easy and simple to publish the data.

- ***Two*** $\star\star$: Achieving *two stars* the user can be able to 1) process the data, 2) aggregate the data, 3)perform calculations, 4) visualise the data and can be exported into another (structured) format. Whereas being data publisher it is still simple to publish.

- ***Three*** $\star\star\star$: Achieving *three stars* helps data user to do all what can be done using *two stars* Web data and additionally one can manipulate the data without any proprietary software package. Similarly as data publisher a converter may be needed to export the data from the proprietary format.

- ***Four*** $\star\star\star\star$:Achieving *four stars* helps data user to do all what can be done using *three stars* Web data and additionally 1) data can be linked using URI's, 2) data can be partly accessed and 3) existing tools and libraries can be reused. On the other hand as data publisher 1) using RDF "Graph" of data can be more effort than tabular (Excel/CSV) or tree (XML/JSON) data, data can be combined safely with other data, 2) data can be merged

[2] `http://sparqles.ai.wu.ac.at/` l.a (10-04-2018)

and combined safely, 3) data can be presented using URIs which is a global scheme for representing data. Being publisher 1) one has fine-granular control over the data items and can optimise their access (load balancing, caching, etc.), 2) data can easily be sliced and diced, 3)URIs must be assigned to data.

- ***Five*** ⋆ ⋆ ⋆ ⋆ ⋆: Achieving *five stars* helps data user to do all what can be done using *four stars* Web data and additionally 1) More data can be discovered while consuming the data, 2) data schema can directly be learned, 3) consumers have to deal with broken data links, and 4) One have to be cautious for linking new data with existing for trust and provenance related issues. On the other hand being data provider or publisher one 1) must have to make data discover-able, 2) can increase the value of the data, 3) can gain the same benefits from the links as the consumers, 4) need to invest resources to link data to existing data over the Web and 5) may face some overhead to repair broken or incorrect links.

From the first star to the last star, the openness of the data is supported by better structure, better interoperability, and thus better reuse through open access, openness of data standards and the interlinking of data as a World Wide Web (WWW) data resource.

4 Fair Data

As proposed by Wilkinson et al., [16] the FAIR principles provide guidelines so that any resource over the Web including data resources, tools, vocabularies and infrastructures should exhibit certain characteristics to ensure the discovery and reuse by anyone. These principles also provide guidelines for data publishing, exploration, sharing, and reuse in both manual as well as automated settings. There have been a number of recent domain specific practices and guidelines for data publishing and management [15,13,6], FAIR guiding principles stand unique in number of ways as these are domain-independent and high-level that can be applied to both metadata as well as data.

In the following we present the elements of FAIR Principles directly taken from Wilkinson et al., [16]. It is worth noting that these elements are related but independent of each other.

The FAIR Guiding Principles [16]:
To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
  => corresponds to the usage of URIs (i.e. ***Four*** ⋆ ⋆ ⋆⋆).

- F2. data are described with rich metadata (defined by R1 below)
  => this certainly requires structured representation (i.e. ***Two*** ⋆⋆) and at best would relate to linkage to other data for context (i.e. ***Five*** ⋆ ⋆ ⋆ ⋆ ⋆).

– F3. metadata clearly and explicitly include the identifier of the data it describes
=> corresponds to the usage of URIs (i.e. ***Four*** ⋆ ⋆ ⋆⋆).

– F4. (meta)data are registered or indexed in a searchable resource
=> this is a statement about technologies making use of the data, which is the consequence of data access over the Web (i.e. ***One*** ⋆)

To be Accessible:

– A1. (meta)data are retrievable by their identifier using a standardised communications protocol
=> corresponds to the usage of URIs (i.e. ***Four*** ⋆ ⋆ ⋆⋆), although other identifiers could be used as well.

– A1.1 the protocol is open, free, and universally implementable
=> corresponds again to URIs and related open data standards (i.e. ***Four*** ⋆ ⋆ ⋆⋆).

– A1.2 the protocol allows for an authentication and authorisation procedure, where necessary
=> the LOD only covers the open access part as the primary requirement (i.e. ***One*** ⋆).

– A2. metadata are accessible, even when the data are no longer available
=> At best, this corresponds to the URIs, but long-term acquisition is not a matter of data standards but relates to IT infrastructures.

To be Interoperable:

– I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
=> this corresponds to the use of "non-proprietary open formats" (i.e. ***Three*** ⋆ ⋆ ⋆).

– I2. (meta)data use vocabularies that follow FAIR principles
=> this statement is self-referential and refers to contextual standardised data (i.e. ***Five*** ⋆ ⋆ ⋆ ⋆ ⋆).

– I3. (meta)data include qualified references to other (meta)data
=> again this refers to contextual standardised data (i.e. ***Five*** ⋆ ⋆ ⋆ ⋆ ⋆).

To be Reusable:

– R1. meta(data) are richly described with a plurality of accurate and relevant attributes

=> this refers to the structured data, but makes a stronger reference to contextual information (i.e. ***Three*** ⋆ ⋆ ⋆).

- R1.1. (meta)data are released with a clear and accessible data usage license
  => a statement about the delivery of (meta)data, which is meant to be open according to LOD (i.e. ***One*** ⋆).

- R1.2. (meta)data are associated with detailed provenance
  => provenance is part of the contextual data (i.e. ***Five*** ⋆ ⋆ ⋆ ⋆ ⋆).

- R1.3. (meta)data meet domain-relevant community standards
  => again, community standards are part of the contextual data and should be openly available (according to LOD, i.e. ***Five*** ⋆ ⋆ ⋆ ⋆ ⋆).

Like LOD five stars these principles are incremental and can be considered in any combination result in achieving the higher degrees of 'FAIRness'. In other words, the better the adoption of FAIR guidelines the higher is the degree of 'FAIRness' claimed.

For each FAIR principle, we give a judgement, which LOD star is best aligned with the FAIR principle, if any applies. As can be seen from the table. 1, more than one LOD principle could be linked to a FAIR principle, but we only give a judgement to the highest LOD 5-star principle.

5 Distinctions and Overlaps

5.1 LOD and non-data assets

The LOD 5 stars scheme is for the Open Data. This scheme provides the guidelines for data providers and publishers in order to make data more accessible, available and reusable. Whereas the FAIR principles can equally be applied to any non-data assets in the same manner as data.

FAIR does not impose any constraint that the data must be openly available and requires that access to the license agreement is made available. The concern can be raised that restricted data will also limit reusability only by the limitation that access is not openly available.

Furthermore, FAIR asks for meta-data to be provided with the data to improve interoperability, which relates to the contextual information from the LOD principles through interoperable data linked as one or several interoperable open data sources. Here, the FAIR data principles envision data about data to improve reusability.

Next, FAIR sees access to the FAIR data as a task which would have to be achieved through a supported IT infrastructure. This is certainly a requirement which would play a more important role on data that is not openly available (non open access data), since the data cannot be openly replicated without license agreement.

Table 1: FAIR principles mapping to LOD 5-star scheme.

	★	★★	★★★	★★★★	★★★★★
F1				X	
F2					X
F3				X	
F4	X				
A1		X	X	X	
A1.1	X	X	X	X	
A1.2	X				
A2					
I1	X	X	X		
I2	X	X			X
I3				X	X
R1			X		X
R1.1	X				
R1.2				X	X
R1.3	X	X			X

Finally, the LOD principles reside on URIs as the key element to achieve openness, interoperability and reuse, whereas FAIR would allow for a wider range of identifiers that would achieve the same purpose. This principle certainly – again – applies mainly to data sources that are more restricted in comparison to Linked Open Data.

5.2 Tool and technology independent

The high-level FAIR Guiding Principles are independent of implementation choices, and do not suggest any specific technology, standard, or implementation-solution. On the other hand, these principles are not, themselves, a standard or a specification. They act as a guide to data publishers as well as consumers to assist them in evaluating whether any specific implementation choices are able to support their digital research artefacts Findable, Accessible, Interoperable, and Reusable [16]. Same is the case with LOD 5-star scheme, which are not themselves a standard or a specification but provides the guideline for choosing the tools and technologies to make linked open data more available, searchable and open.

Acknowledgement

This research has been supported in part by Science Foundation Ireland under Grant Number SFI/12/RC/2289.

References

1. Berners-Lee, T.: Is your linked open data 5 star. Repéré à https://www.w3.org/DesignIssues/LinkedData.html (2010)
2. Berners-Lee, T., Fischetti, M., Foreword By-Dertouzos, M.L.: Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor. HarperInformation (2000)
3. Bourne, P.E., Clark, T.W., Dale, R., de Waard, A., Herman, I., Hovy, E.H., Shotton, D.: Improving The Future of Research Communications and e-Scholarship (Dagstuhl Perspectives Workshop 11331). Dagstuhl Manifestos 1(1), 41–60 (2012), http://drops.dagstuhl.de/opus/volltexte/2012/3445
4. Cox, S., Yu, J.: Oznome 5-star tool: A rating system for making data fair and trustable. In: Proceedings of the 2018 eResearch Australasia Conference (2017)
5. Hasnain, A., Fox, R., Decker, S., Deus, H.F.: Cataloguing and linking life sciences LOD Cloud. In: 1st International Workshop on Ontology Engineering in a Data-driven World collocated with EKAW12 (2012)
6. Hasnain, A., Kamdar, M.R., Hasapis, P., Zeginis, D., Warren Jr, C.N., et al.: Linked Biomedical Dataspace: Lessons Learned integrating Data for Drug Discovery. In: International Semantic Web Conference (In-Use Track), October 2014 (2014)
7. Hasnain, A., Mehmood, Q., e Zainab, S.S., Decker, S.: A provenance assisted roadmap for life sciences linked open data cloud. In: Knowledge Engineering and Semantic Web, pp. 72–86. Springer (2015)
8. Hasnain, A., Mehmood, Q., e Zainab, S.S., Hogan, A.: Sportal: Profiling the content of public sparql endpoints. International Journal on Semantic Web and Information Systems (IJSWIS) 12(3), 134–163 (2016), http://www.igi-global.com/article/sportal/160175
9. Hasnain, A., Mehmood, Q., e Zainab, S.S., Hogan, A.: Sportal: Searching for public sparql endpoints. In: International Semantic Web Conference (Posters & Demos) (2016)
10. Hasnain, A., e Zainab, S.S., Kamdar, M.R., Mehmood, Q., Warren Jr, C.N., Fatimah, Q.A., Deus, H.F., Mehdi, M., Decker, S.: A roadmap for navigating the life sciences linked open data cloud. In: Semantic Technology, pp. 97–112. Springer (2014)
11. Hasnain, S.M.A.: Cataloguing and linking publicly available biomedical SPARQL endpoints for federation-addressing aPosteriori data integration. Ph.D. thesis (2017)
12. Saleem, M., Hasnain, A., Ngomo, A.C.N.: Largerdfbench: a billion triples benchmark for sparql endpoint federation. Journal of Web Semantics (2018)
13. Sandve, G.K., Nekrutenko, A., Taylor, J., Hovig, E.: Ten simple rules for reproducible computational research. PLoS computational biology 9(10), e1003285 (2013)
14. Vandenbussche, P.Y., Umbrich, J., Matteis, L., Hogan, A., Buil-Aranda, C.: Sparqles: Monitoring public sparql endpoints. Semantic Web 8(6), 1049–1065 (2017)

15. White, E.P., Baldridge, E., Brym, Z.T., Locey, K.J., McGlinn, D.J., Supp, S.R.: Nine simple ways to make it easier to (re) use your data. PeerJ PrePrints (2013)
16. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data 3 (2016)
17. Yamamoto, Y., Yamaguchi, A., Splendiani, A.: Yummydata: providing high-quality open life science data. Database 2018 (2018)