SCIENTIFIC DATA

OPEN

COMMENT

FAIR adoption, assessment and challenges at UniProt

Leyla Garcia[1], Jerven Bolleman[2], Sebastien Gehant[2], Nicole Redaschi[2], Maria Martin[1] & UniProt Consortium*

Received: 28 May 2019
Accepted: 23 August 2019
Published online: 20 September 2019

UniProt continues to support the ongoing process of making scientific data FAIR. Here we contribute to this process with a FAIRness assessment of our UniProtKB dataset followed by a critical reflection on the challenges and future directions of the adoption and validation of the FAIR principles and metrics.

Data management and stewardship plans are nowadays essential to ensure the long-term sustainability of digital assets. The Findable, Accessible, Interoperable and Reusable (FAIR) principles[1], first described in 2016, provide a framework defining the minimum elements required for good data management, making it easier for data providers to offer support for data driven knowledge discovery and innovation. Some of the main points of the FAIR principles address identification, licensing and data longevity policies.

Adopting the FAIR principles has proven to be a complex task that involves not only knowledge of your own data, but also awareness of metadata, schemata, protocols, policies, and community agreements. Another challenge lies in the vagueness of the original FAIR principles which offer a foundation layer for data management, but do not formally define how to fulfil the different elements under consideration. As a consequence, data providers may choose among a diversity of possible implementations making it difficult to critically assess the FAIRness of any resource. In order to overcome such limitations, a set of exemplar metrics were published in 2018[2] and later complemented by a FAIR maturity framework[3].

Although the importance of FAIR has been recognized widely by the research community via initiatives such as GO-FAIR (https://www.go-fair.org/) as well as a series of workshops to assess the FAIRness of current ELIXIR Core Data Resources (https://www.elixir-europe.org/platforms/data/fairness-core-resources), the adoption of the principles is still an ongoing process. Here we report our contribution to the process of FAIR adoption in the form of a FAIRness assessment on the Universal Protein Resource (UniProt)[4]. UniProt is a comprehensive resource for protein sequence and annotation data; it provides three main datasets: the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef) and the UniProt Archive (UniParc), all of them released every four weeks. UniProtKB is a central hub for the collection of functional information on proteins including accurate, consistent and rich annotation. UniRef provides clustered sets of sequences from UniProtKB and selected UniParc records. UniParc is a non-redundant dataset containing most of the publicly available protein sequences. With this FAIRness assessment, we aim to share our experience and the challenges we met with other resource providers and FAIR initiatives, so our experience can be used to further refine the FAIR principles and metrics.

Our FAIRness Assessment Journey for UniProt

A FAIRness assessment for a large resource such as UniProt is not straight forward. UniProt data are published via a website (https://www.uniprot.org/uniprot) and distributed in multiple serialization formats, including a custom text format, XML, RDF/XML and FASTA. In addition, we also provide Application Programming Interfaces (API)s and File Transfer Protocol (FTP) downloads. The first question that we encountered during our FAIR assessment concerned this range of different distribution formats. Should all distributions be assessed as one or separately? Other resources that also support multiple serialization formats could face the same question when assessing their resources against the FAIR principles and metrics. In order to overcome these difficulties, ELIXIR Europe has supported a series of workshops to assess the FAIRness status of ELIXIR Core Data Resources (https://

[1]European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. [2]SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211, Geneva 4, Switzerland. *A comprehensive list of consortium members appears at the end of the paper. Correspondence and requests for materials should be addressed to L.G. (email: ljgarcia@ebi.ac.uk)

www.elixir-europe.org/platforms/data/fairness-core-resources). The resulting recommendation from one of these workshops was to use the FAIRest distribution. In the case of UniProt, this is its RDF/XML representation as this is based on well-structured vocabularies, i.e., schemata. This does, however raise the question: can a resource really be FAIR or just have FAIR representations? We do not know the answer yet, but more will be learnt as additional resources move to become FAIR with supporting FAIRness assessment.

We decided to use the UniProtKB dataset to assess the FAIRness of UniProt data because it is the most complex and most widely used of the three main UniProt datasets. We have followed the exemplar FAIR metrics[2] together with supplementary information providing questions and assessments for other resources as reported by the FAIR maturity framework[3]. This assessment has been undertaken using UniProt release 2019_02. From one release to another, metadata such as dates and version together with the content itself are updated, but aspects such as identification schemata, access protocols and license usually remain the same. In the Online-only Table 1 we present our FAIR assessment results together with some supporting information.

Based on our assessment, UniProt is almost completely FAIR, with some remaining issues regarding the requirement for certification provided by a recognized authority. We cannot yet address these issues because it is currently unclear what a recognized authority, either FAIR or community based, would be for proteins.

A Word on Identifiers, Metadata, and Data

The FAIR principles were designed for digital resources, their metadata and data. In order to relate a digital resource to their data content, there needs to be an explicit link between them. UniProt has an identifier as a dataset as a whole, "http://purl.uniprot.org/void#UniProtDataset". Additionally, each set of data in UniProt, which we define as each UniProtKB entry in our assessment, also has an identifier, for example "https://purl.uniprot.org/uniprot/P05067". Following the FAIR principles, all identifiers should be included in the respective metadata. From the dataset it should be possible to get to the content, i.e., UniProtKB entries in our case, or vice versa; whatever the chosen direction, dataset and content should be linked to each other. In the case of big datasets such as UniProt, the list of the entries contained in the dataset becomes too long to be included in the dataset metadata. A feasible alternative is to include a link from the entry to the dataset. If needed, a complete list of the dataset entries could be compiled by programmatic means, such as a SPARQL query designed to retrieve all entries included in dataset version 2019_02. Introducing a pattern-like link as part of the resource metadata would make it easier to reach its content. For validation purposes, an exemplar content identifier could also be included. This is a case that could be considered in the FAIR metrics. In the case of UniProtKB entries, such an identification pattern for content identifiers does exists and is documented in the Help pages (https://www.uniprot.org/help/accession_numbers).

In addition to the described link between resources and content, it is also important to take into account differences across multiple representations of a same dataset. In UniProt, the concept of an entry makes sense for our XML and custom text format, but it is hard to apply to the RDF world where each statement is an independent entity. For example, there are over 140 million UniProtKB entries in the 2019_02 dataset, but the corresponding RDF distribution also includes statements about many more International Nucleotide Sequence Database Collaboration (INSDC) "entries", as well as over one billion other linked database "entries". We also have to consider that most of our users do not want to retrieve what we consider to be a full dataset and will compose their own "subsets" via website or API queries, and we have therefore chosen to make each entry independently accessible.

Finally, the distinction between metadata and data is in many ways an arbitrary one. For some of our users the evidence for our assertions, e.g., publications, are metadata, while for other users they are critical data. Some serialization formats, especially those designed to be used by software tools, e.g. FASTA or GFF, make it impractical to include all data and metadata.

Challenges and Evolution of FAIRness Assessments

We recognized the complexity that a large resource like UniProt poses for a FAIRness assessment. Even for smaller datasets FAIRness assessments are not a straightforward process. The current exemplar metrics, together with their question set are definitely a step forward in facilitating the FAIRification of resources; nonetheless, the process is still manual and requires human verification of the answers. Some of the questions such as those about schemata behind the identifiers and protocols, relate to third-party URLs, which are not necessarily in a machine-readable format. Information about HTTP or HTTPS can be found in Wikipedia, but would that be the correct URL for a FAIRness assessment? We do not know the answer and the metrics and questions do not help here. We mimicked the assessment examples provided as supplementary material at the GitHub FAIR metrics repository, as this seemed to be the simplest approach at this time.

The pilot project FAIRshake[5] aims to make manual assessments easier. It presents users with a set of questions that are similar to those accompanying the exemplar metrics. The assessment process is still manual, based on questions and IRIs, and therefore presents the assessor with similar issues as does the question set accompanying the exemplar metrics. Rather than relying on manual assessments, the FAIR community should aim to create a semi or even fully automated validator to make assessments easier and comparable. Such a validator could, for instance, take account of the third-party URLs mentioned at the beginning of this section.

The FAIR principles and metrics are still evolving. They are gaining a momentum that should push digital resources to face the FAIR challenges and, by doing so, improve science. Communities will play an important role to make this a reality and the FAIR principles recognized this, for instance the principle F2 refers to rich metadata, R1 mentions a plurality of relevant attributes and R1.3 talks about community standards. Any FAIR validator should therefore be complemented with community-based validators. There are different accepted standards for datasets, e.g., DCAT (https://www.w3.org/TR/vocab-dcat/), EOSC-EDMI (https://eosc-edmi.github.io/) and Bioschemas[6] (http://bioschemas.org/). For RDF distributions there is the external FAIR validator

at YummyData[7] (http://yummydata.org/) which strives to generate a computable FAIR metric. The data that we provide to YummyData are also used to improve our user documentation for the UniProt SPARQL endpoint at sparql.uniprot.org. This shows how being FAIR can also benefit the resource providers themselves.

While the FAIR principles and metrics cover a minimum of elements such as identifiers, license and provenance, community standards could go a step further by requiring additional metadata, thus improving interoperability and reusability. Despite their importance, data catalogs and datasets are not the only digital resources in existence. We expect that additional FAIR communities will emerge to adapt the existing principles to other digital resources such as training materials, software and services. The principles will then be tested outside their initial scope and adapted to add further exciting chapters to this FAIR tale.

References

1. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018, https://doi.org/10.1038/sdata.2016.18 (2016).
2. Wilkinson, M. D. *et al.* A design framework and exemplar metrics for FAIRness. *Sci. Data* **5**, 180118, https://doi.org/10.1038/sdata.2018.118 (2018).
3. Wilkinson, M. D. *et al.* Evaluating FAIR Maturity Through a Scalable, Automated, Community-Governed Framework. *Preprint at*, https://doi.org/10.1101/649202 (2019).
4. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515, https://doi.org/10.1093/nar/gky1049 (2018).
5. Clarke, D. J. B. *et al.* FAIRshake: toolkit to evaluate the findability, accessibility, interoperability, and reusability of research digital resources. *Preprint at*, https://doi.org/10.1101/657676 (2019).
6. Gray, A. J. G., Goble, C. A. & Jimenez, R. C. Bioschemas: From Potato Salad to Protein Annotation. In *International Semantic Web Conference (Posters, Demos & Industry Tracks)* (Vienna, Austria, 2017).
7. Yamamoto, Y., Yamaguchi, A. & Splendiani, A. YummyData: providing high-quality open life science data. *Database* **2018**, bay022, https://doi.org/10.1093/database/bay022 (2018).

Acknowledgements

We acknowledge the organizers and participants of the Nettab 2018 workshop for their comments during the poster presentation of a preliminary assessment of the UniProt datasets. Likewise, we acknowledge the organizers and participants of the EMBL-EBI FAIRness assessment for data core resources workshop on May 2018 for their discussion regarding FAIR metrics and questions. Finally, we want to acknowledge our funders. UniProt is supported by the National Eye Institute (NEI), National Human Genome Research Institute (NHGRI), National Heart, Lung, and Blood Institute (NHLBI), National Institute on Aging (NIA), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of General Medical Sciences (NIGMS), and National Institute of Mental Health (NIMH) of the National Institutes of Health (NIH) under grant U24HG007822. Additional support for the EMBL-EBI's involvement in UniProt comes from European Molecular Biology Laboratory (EMBL), Alzheimer's Research UK (ARUK-NAS2017A-1), the British Heart Foundation (BHF) (RG/13/5/30112), the Parkinson's Disease United Kingdom (PDUK) GO grant G-1307, and the NIH GO grant U41HG02273. UniProt activities at the SIB are additionally supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation SERI. PIR's UniProt activities are also supported by the NIH grants R01GM080646, G08LM010720, and P20GM103446, and the National Science Foundation (NSF) grant DBI-1062520. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Consortia
UniProt Consortium

Alex Bateman[1], Michele Magrane[1], Maria Martin[1], Sandra Orchard[1], Shriya Raj[1], Shadab Ahmad[1], Emanuele Alpi[1], Emily Bowler[1], Ramona Britto[1], Borisas Bursteinas[1], Hema Bye-A-Jee[1], Tunca Dogan[1], Leyla Garcia[1], Penelope Garmiri[1], George Georghiou[1], Leonardo Gonzales[1], Emma Hatton-Ellis[1], Alexandr Ignatchenko[1], Giuseppe Insana[1], Rizwan Ishtiaq[1], Vishal Joshi[1], Dushyanth Jyothi[1], Jie Luo[1], Yvonne Lussi[1], Alistair MacDougall[1], Mahdi Mahmoudy[1], Andrew Nightingale[1], Carla Oliveira[1], Joseph Onwubiko[1], Vivek Poddar[1], Sangya Pundir[1], Guoying Qi[1], Ahmet Rifaioglu[1], Daniel Rice[1], Rabie Saidi[1], Elena Speretta[1], Edward Turner[1], Nidhi Tyagi[1], Preethi Vasudev[1], Vladimir Volynkin[1], Kate Warner[1], Xavier Watkins[1], Rossana Zaru[1], Hermann Zellner[1], Alan Bridge[2], Lionel Breuza[2], Elisabeth Coudert[2], Damien Lieberherr[2], Ivo Pedruzzi[2], Sylvain Poux[2], Manuela Pruess[2], Nicole Redaschi[2], Lucila Aimo[2], Ghislaine Argoud-Puy[2], Andrea Auchincloss[2], Kristian Axelsen[2], Parit Bansal[2], Delphine Baratin[2], Teresa Batista Neto[2], Marie-Claude Blatter[2], Jerven Bolleman[2], Emmanuel Boutet[2], Cristina Casals-Casas[2], Beatrice Cuche[2], Edouard De Castro[2], Anne Estreicher[2], Livia Famiglietti[2], Marc Feuermann[2], Elisabeth Gasteiger[2], Sebastien Gehant[2], Vivienne Gerritsen[2], Arnaud Gos[2], Nadine Gruaz[2], Ursula Hinz[2], Chantal Hulo[2], Nevila Hyka-Nouspikel[2], Florence Jungo[2], Arnaud Kerhornou[2], Philippe Lemercier[2], Thierry Lombardot[2], Patrick Masson[2], Anne Morgat[2], Sandrine Pilbout[2], Monica Pozzato[2], Catherine Rivoire[2], Christian Sigrist[2], Shyamala Sundaram[2], Cathy Wu[3,4], Cecilia Arighi[3,4], Hongzhan Huang[3,4], Peter McGarvey[3,4], Darren Natale[3,4], Leslie Arminski[3,4], Chuming Chen[3,4], Yongxing Chen[3,4], John Garavelli[3,4], Kati Laiho[3,4], Karen Ross[3,4], C. R. Vinayaka[3,4], Qinghua Wang[3,4], Yuki Wang[3,4], Lai-Su Yeh[3,4] & Jian Zhang[3,4]

[3]Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven Street, NW, Suite, 1200, 20007, USA. [4]Protein Information Resource, University of Delaware, 15 Innovation Way, Suite 205, Newark, DE, 19711, USA.