Mapping subjects and domains across the humanities and natural sciences in FAIRsharing

Allyson L. Lister[1 [0000−0002−7702−4495]], Peter McQuilton[1 [0000−0003−2687−1982]], Milo Thurston[1 [0000−0002−6468−9260]], and Susanna-Assunta Sansone[1 [0000−0001−5306−5690]]

Oxford e-Research Centre (OeRC), Department of Engineering Science, University of Oxford, Oxford, UK contact@fairsharing.org
https://www.oerc.ox.ac.uk/

Abstract. FAIRsharing is a manually-curated registry of research data standards, repositories and data policies across all humanities and natural science disciplines. Each FAIRsharing record contains over 40 metadata properties that provide a rich description of each resource. Storing over 2,700 records, FAIRsharing is accessed via a web portal, API and other visualisation tools. In order to provide accurate, structured searching, ontologies are used to populate a number of the FAIRsharing metadata fields. Most recently, two application ontologies have been built from over 2000 free-text user-generated tags. These tags were classified as subjects/research areas (SRAO - Subject Resource Application Ontology) or as research domains (DRAO - Domain Resource Application Ontology). To ensure the required breadth of scope and high level of interoperability, and to limit redundancy with other community efforts, these application ontologies are based on a number of well-supported and adopted ontologies. FAIRsharing promotes FAIR and open science, with both ontologies open and freely available via GitHub (SRAO, DRAO).

Keywords: FAIR Data · Application Ontologies · Data Sharing.

1 Introduction

As part of the FAIRsharing [4] curation pipeline, curators and maintainers annotate records with subjects and knowledge domains. Initially, approximately 2000 free text tags were stored as a flat list, creating usability issues and limiting search functionality. Definitions, synonyms, a hierarchical structure, and richer semantics were required. Many publicly-available vocabularies could provide part of what was needed, but none had the breadth of coverage required across all research areas so, to provide the necessary semantics and structure, two application ontologies (AOs), SRAO (Subject Resource Application Ontology) and DRAO (Domain Resource Application Ontology), were created.

Application ontologies (AO) are intended as low-maintenance solutions, importing only the terms and hierarchies required by a project. This drastically reduces the size of the resulting ontology and increases interoperability with pre-existing vocabularies. Within FAIRsharing, automated development and build

procedures together with simple manual curation has produced AOs that are both large in scope and small in build complexity.

2 Results

SRAO is a high-level subject hierarchy spanning all research areas and based on 7 ontologies and classification systems. SRAO's hierarchy is manually curated and, where possible, aligned with external ontologies. Definitions and/or synonyms from these ontologies are automatically imported into SRAO via Ontofox [5] to supplement the existing subject annotation.

DRAO provides fine-grained domain tags from over 50 external ontologies, and is primarily built using automated procedures. Importing of the required terms and their hierarchies was performed using Ontofox via a MIREOT [1]-compliant process. After import, FAIRsharing-specific annotation was added to the classes using ROBOT [2]. Although a number of biomedical application ontologies (e.g. EFO [3]) have been created in this manner, we believe DRAO maps more reference ontologies than any similar effort.

Both SRAO and DRAO are used by FAIRsharing curators and our user community to annotate FAIRsharing records and perform searches across the site. Both SRAO and DRAO are open, community-driven AOs and we hope others will find them appropriate for use in their work. SRAO and DRAO are available for general use via GitHub (SRAO, DRAO) with a CC BY-SA 4.0 licence. Release files are created using ROBOT. Improvements to the ontologies can be suggested via their GitHub trackers and the ontologies are licensed for use in other projects as appropriate.

References

1. Courtot, M., Gibson, F., Lister, A.L., Malone, J., Schober, D., Brinkman, R.R., Ruttenberg, A.: Mireot: The minimum information to reference an external ontology term. Appl. Ontol. **6**(1), 23–33 (Jan 2011), http://dl.acm.org/citation.cfm?id=1971674.1971680
2. Jackson, R.C., Balhoff, J.P., Douglass, E., Harris, N.L., Mungall, C.J., Overton, J.A.: Robot: A tool for automating ontology workflows. BMC Bioinformatics **20**(1) (7 2019). https://doi.org/10.1186/s12859-019-3002-3
3. Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., Parkinson, H.: Modeling sample variables with an Experimental Factor Ontology. Bioinformatics **26**(8), 1112–1118 (03 2010). https://doi.org/10.1093/bioinformatics/btq099, https://doi.org/10.1093/bioinformatics/btq099
4. Sansone, S.A., , McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A.L., Thurston, M.: FAIRsharing as a community approach to standards, repositories and policies. Nature Biotechnology **37**(4), 358–367 (apr 2019). https://doi.org/10.1038/s41587-019-0080-8, https://doi.org/10.1038%2Fs41587-019-0080-8
5. Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A., He, Y.: Ontofox: web-based support for ontology reuse. In: BMC Research Notes (2010)