FAIRsharing: working with and for the community to describe and link data standards, repositories and policies

Susanna-Assunta Sansone, Peter McQuilton, Philippe Rocca-Serra, Alejandra Gonzalez-Beltran, Massimiliano Izzo, Allyson Lister, Milo Thurston and the FAIRsharing community*.

Oxford e-Research Centre, Department of Engineering Science, University of Oxford, Oxford, OX1 3QG, UK.

**Corresponding author:* Susanna-Assunta Sansone susanna-assunta.sansone@oerc.ox.ac.uk and contact@fairsharing.org

Community-developed standards, such as those for the identification and reporting of data, underpin reproducible and reusable research. The number of community-driven efforts has been on the rise since the early 2000s, their uptake, however, is slow and uneven. Analyzing 70 journals and publishers data policies, we find that these recommend databases and repositories 37 times more often than standards. When a reporting standard is recommended by a publisher, it is more likely to be a minimal reporting guideline than a model, format or ontology even if the latter are the machine-readable standards that underpin the utility of databases and repositories. Here, we evaluate the standards landscape, focusing on those for reporting data and metadata, and their implementation by databases and repositories; we also propose key performance indicators, and highlight the importance of developing open linked data models that instantiate these community standards. Lastly, we launch a call to action highlighting the role producers and consumers of standards and repositories must play to maximize the visibility and adoption of these resources.

Assessing and addressing community needs

Community-developed norms and specifications, such as those on citation (1), identification (2) and metadata reporting (e.g. 3), are designed to assist the virtuous research life cycle, from collection to annotation, through preservation and publication, to subsequent sharing and reuse of digital artifacts (e.g. data, articles, software, models, workflows). Commonly referred to as community standards, these norms and

specifications enable reproducible research, reduce duplication of effort, aid scholarly publishing, and drive both discovery and the evolution of scientific practice.

It is exciting therefore to see that journals (e.g. 4) pledges to help mandate the use of standards and repositories as a condition of publication, focusing on those that are clearly established and maintained by the research community. We applaud those stakeholders who take concrete steps to promote data sharing and open science, rather than just advocate for it. However, we offer a word of caution: arbitrary decisions that promote one resource over another can be worse than empty rhetoric. There is an urgent need for objective indicators to help stakeholders make informed decisions, especially as to which standards and repositories to use or endorse. But first and foremost, we need to paint an accurate landscape of the evolving constellation of heterogeneous options available.

There are thousands of community-developed standards across all disciplines, some of which have been created and/or implemented by several thousand data repositories. As any other digital object, standards and repositories have a life cycle that encompasses formulation, development and maintenance (5); their status in this cycle (i.e. are they still in development, ready to use, or deprecated/superseded) may vary depending on how active their developing community is. For the consumers of these resources, it can be difficult to know which are the most relevant standards for a specific discipline or need, or at what level of maintenance they are, and which repositories implement them. Conversely, for the producers of standards and repositions it is important that their resources are discoverable by prospective users both within and outside their direct discipline, to foster collaboration and reduce the potential for unnecessary reinvention.

Mapping the landscape

Working with and for the producers and consumers of these standards for over 17 years has provided us with invaluable insights into their life cycle along with a network of international collaborators; all essential elements to tackle this challenge. We have developed FAIRsharing (https://fairsharing.org), a curated, informative and educational resource, describing and interlinking standards, databases, repositories, and data policies. Some readers may recognize FAIRsharing by its former name of BioSharing (6,7), launched in 2011 and born from the MIBBI portal released in 2008 (3).

As of December 2017, FAIRsharing has over 2175 records mainly relevant to the life, agricultural, environmental, biomedical and health sciences, and is progressively expanding to other disciplines due to community demand. FAIRsharing brings the producers and consumers of standards closer together and has a growing list of adopters including standardization groups, databases and repositories developers, research data management support initiatives, service providers, curators, data managers, librarians, journal publishers and policy makers (https://fairsharing.org/communities). Using community participation, the

FAIRsharing team accurately describes community-driven standards, such as minimum reporting guidelines (or checklists), models/formats and terminologies (such as taxonomies or ontologies) - ranging from generic and multi-disciplinary, to those from specific disciplines; it makes them discoverable and monitors their evolution, implementation in databases and repositories, and recommendation in journal and funder data policies. An example of a community input that has helped to shape FAIRsharing is the survey (8) run in 2016, which gathered 533 responses from a variety of users and stakeholders on which features and descriptors they needed to make informed decision as to which standards and repositories to use or endorse.

Contributing to the FAIR ecosystem

There is no better name for a resource that works with and for the community to collect the necessary information to ensure standards, repositories and data policies are Findable (e.g., by providing functionalities to register, claim, maintain, inter-link, search and discover them), Accessible (e.g., identifying their level of openness) and encourage they become Interoperable and Reusable, according to the FAIR principles (9). With the goal of being an interoperable component in the ecosystem of other resources and services, we are in the process of minting Digital Object Identifiers (DOIs) to provide a persistent and unique identifier for referencing our records; also we work with the FAIR Metrics group (http://fairmetrics.org; 10) to develop measurable indicators - which we will implement in the FAIRsharing registry progressively - to guide producers to assess the level of FAIRness of their resources.

We also collaborate with several research and infrastructure programmes, which are generic and across disciplinary, such as the Global and Open (GOFAIR; http://go-fair.org) and the European Open Science Cloud (EOSC; https://eoscpilot.eu), and discipline specific such as ELIXIR (https://www.elixir-europe.org), the US National Institutes of Health (NIH) Big Data to Knowledge Initiative and the FAIR Data Commons (https://commonfund.nih.gov/bd2k/commons). Among others, FAIRsharing features in the upcoming reports by Science Europe (https://www.scienceeurope.org) on "Discipline-specific Research Data Management", and by JISC (https://www.jisc.ac.uk) on “FAIR in practice”.

To further expand our community engagement work, FAIRsharing also operates as an open working group under two global community initiatives, such as Force11 (https://www.force11.org) and the Research Data Alliance (RDA; https://www.rd-alliance.org). First and foremost, the working group is finalizing a set of recommendation to guide consumers and producers of standards and repositories to select and describe them, or recommend them in data policies, but it is also collaborating with other RDA and Force11 groups to define a common framework for research data policy, and identify criteria and develop tools for the selection of standards, databases and repositories, e.g., when creating data management plans.

We say we need standards, but do we use them?

The scientific community, including funders and publishers, endorses the concept that common data and metadata standards underpin data reproducibility, ensuring that the relevant elements of a dataset are reported and shared consistently and meaningfully. But navigating through the many standards available can be discouraging and often unappealing for prospective users. Bound by a particular discipline or domain, reporting standards are fragmented, with gaps and duplications. Understanding how they work or how to comply with them takes time and effort.

FAIRsharing plays its part in providing a snapshot of the standards landscape. **Figure 1** and **Table 1** provide a manually curated view on the status quo. However, be aware that this landscape is dynamic and will continue to evolve as we engage with more communities to verify the information we house, add new standards, track their life-cycle status and usage, and link out to examples and training material, where available.

Figure 1. The number of reporting guidelines, models/formats and terminologies, as of December 2017; 575 of which are specific to the life, agricultural, environmental, biomedical and health sciences, and 30 are generic and multi-disciplinary. Indicators show the status in their life cycle: 'Ready' for use, 'In Development', 'Uncertain' when any attempt to reach out to the developing community has failed, and 'Deprecated' when available the reason is detailed in the deprecated record. https://doi.org/10.6084/m9.figshare.5303188.

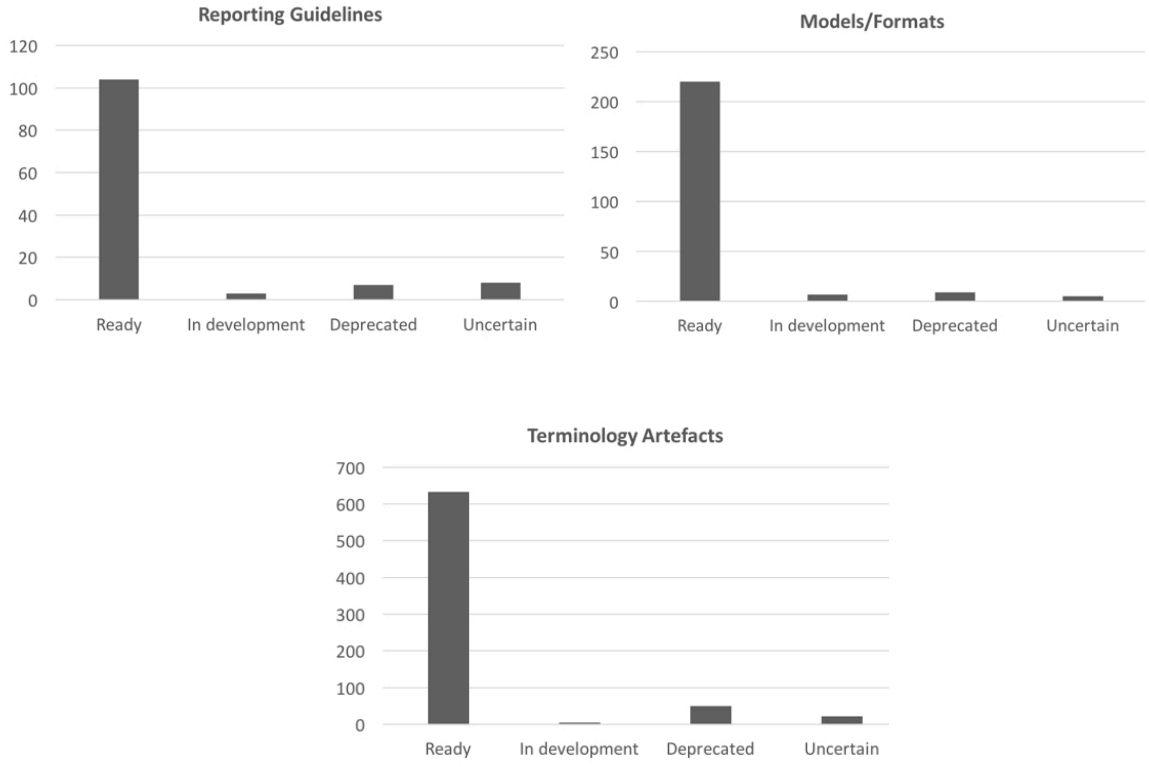


Table 1. The top ten standards more accessed during 2016. This rank, however, shows no direct correlation with their level of adoption (by journals and databases' data policies, databases and repositories), and it probably reflects the activity of their community, in the case those in development, and their popularity within their direct domain. https://doi.org/10.6084/m9.figshare.5303416.v1

Name	Type	Page views	Life cycle status	Number of journals and publishers policies recommending it	Number of databases and repositories implementing it
1. CDISC ADaM https://fairsharing.org/bsg-s000001	Model/format	343	Ready	0	0
2. MIAME https://fairsharing.org/bsg-s000177	Reporting guideline	295	Ready	2	4
3. MIAPPE https://fairsharing.org/bsg-s000543	Reporting guideline	214	Ready	0	2
4. MINSEQE https://fairsharing.org/bsg-s000174	Reporting guideline	210	Ready	1	3
5. MIxS- MIGS/MIMS https://fairsharing.org/bsg-s000161	Reporting guideline	170	Ready	0	2
6. MIxS https://fairsharing.org/bsg-s000518	Reporting guideline	158	Ready	3	3

7. MIAPTE https://fairsharing.org/bsg-s000671	Reporting guideline	145	In Development	n/a	n/a
8. BioPAX https://fairsharing.org/bsg-s000038	Model/format	142	Ready	0	2
9. ISA-Tab https://fairsharing.org/bsg-s000078	Model/format	134	Ready	3	8
10. AnIML https://fairsharing.org/bsg-s000545	Model/format	129	In Development	n/a	n/a

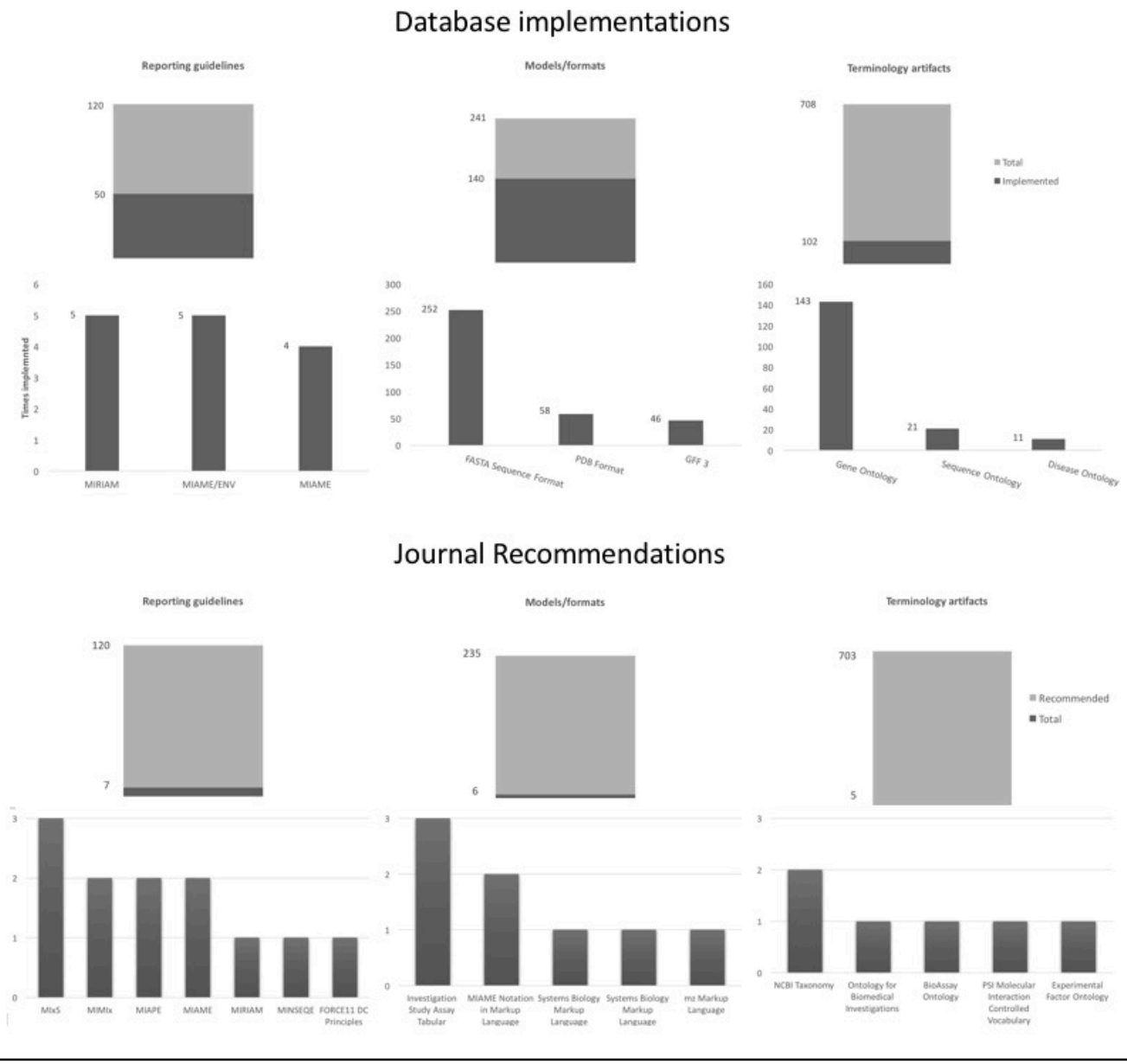
Activating the decision-making chain is an essential step. When a standard is mature and appropriate standard-compliant systems become available, such as databases and repositories, these must then be channelled to the relevant stakeholder community, who in turn must recommend them (e.g. in data policies) or use them (e.g. to define a data management plan) to facilitate a high-quality research cycle.

To understand how journals and publishers select the resources to recommend, we have worked closely with the editors of eight journals/publishers, whose data policies are quite well developed. The resources that EMBO Press, F1000Research, Oxford University Press' *GigaScience*, PLOS, Elsevier, Wellcome Trust' *Wellcome Open Research* and Springer Nature's BioMed Central and *Scientific Data* recommend are explorable at https://fairsharing.org/recommendations. The data policies of these eight journals/publishers recommend a total 18 standards (7 reporting guidelines, 6 models/formats, 5 terminologies), and 185 databases and repositories. However, there are additional 180 standards that should be explicitly mentioned in these data policies because they are implemented by the recommended databases and repositories.

Analyzing the total 70 journals and publishers data policies curated in FAIRsharing, as of December 2017 (https://fairsharing.org/policies/?q=&selected_facets=subtype_exact:Journal), we find that 56 mention one or more specific standards; see **Figure 2**. When standards are recommended by data policies, the minimal reporting guidelines are recommended 1.6 times more than terminology artifacts and models/formats, even if

the latter two are heavily implemented by databases and repositories. In general, databases and repositories are 37 times more recommended than models/formats; even when a recommended database or repository implements specific models/formats, the latter are not explicitly mentioned by the data policies.

Figure 2. The total number of reporting guidelines, models/formats and terminologies, as of December 2017, and the top three from each type, implemented by databases and repositories and recommended by journals and publishers' data policies. https://doi.org/10.6084/m9.figshare.5303206.v1



It would not be FAIR if standards were not executable

The under-representation of recommended terminology artifacts and models/formats is of particular concern. Minimal reporting guidelines are intended for human consumption and are usually narrative in form and therefore prone to ambiguities, making compliance and validation difficult and approximated. Many of these guidelines however, already come with (or lead to the development of) associated models/formats and terminology artifacts, which instead are created for machine consumption.

The latter are essential to the FAIR principles, which put a specific emphasis on enhancing the ability of machines to automatically discover and use data. In particular, the computability of metadata standards is core to the development of metrics of FAIRness to measure the level of compliance of a given dataset against the relevant metadata descriptors. These machine-readable standards provide the necessary quantitative and verifiable measures of the degree by which data meets these reporting guidelines. The latter, on their own would just be statements of unverifiable good intentions of compliance to given metadata standards.

Help us to help you

To promote the use of standards, databases and repositories and paint an accurate picture of their relationships, four main stakeholders can play catalytic roles.

- **Developers and curators of standards, databases, repositories.** FAIRsharing helps you to make your resources more discoverable, gain increased exposure and credit outside of your immediate community, promoting adoption (learn how to add your resource to FAIRsharing, or claim it, at https://fairsharing.org/new).
    - As the representative of a community standardization initiative, you are best placed to describe the status of your standards and track their evolution (creating an individual record e.g., the DDI standard for social, behavioral, economic, and health data: https://fairsharing.org/bsg-s000605; or grouping several records in a collection e.g., the HUPO PSI standards for proteomics and interactomics data: https://fairsharing.org/collection/HUPOPSI). If you strive for FAIR data, then you need to ensure you also deliver linked data models that allow the publishing and connecting of structured data on the web.
    - Similarly, as representative of a database or repository, you are uniquely placed to describe your resource, and to declare the standards you implement (e.g., the ICPSR archive of behavioral and social science research data that uses the DDI standard: https://fairsharing.org/biodbcore-000936; or the Reactome knowledge base, https://fairsharing.org/biodbcore-000329, which uses several standards in the COMBINE

collection for computational models in biology networks: https://fairsharing.org/collection/ComputationalModelingCOMBINE).

- **Journal editors, publishers or an organization with a data policy.** FAIRsharing helps you maintain an interrelated list of standards, databases and repositories, grouping those you want to recommend to your users (e.g., see examples of recommendations created by eight main publishers and journals: https://fairsharing.org/recommendations; and the record of the UniProt Knowledgebase as an example of a highly recommended repository https://fairsharing.org/biodbcore-000544). You can learn more about standards, especially those implemented by databases and repositories you already recommend, as you should explicitly mention these standards in your policy. As we continue to map the landscape, you can also revise your selections over time, recommending further resources with more confidence.
- **Trainers and educators.** FAIRsharing provides you with a base to create or enrich training material on the role and use of standards in databases and repositories to enable research data management and reproducibility. Enhancing both the capability and skills of those involved in producing, managing, serving, curating, preserving, publishing or regulating data (and other digital objects) is a vital step to reduce the knowledge gap on standards that is currently found in the research community.
- **Funding agencies**. FAIRsharing helps you select resources to recommend in your data policy, or that awardees should consider when writing their data management plan. If we are to make FAIR data a reality, you should recognize standards as digital objects in their own right, with their associated research, development and educational activities (5). New funding frameworks need to be created to provide catalytic support for this techno-social activities, in specific domains, within and across disciplines to enhance interoperability of data.

No more hollow promises, it is time for hard outcomes. It won't be FAIR if was easy. Roll up our sleeves, whatever the color of your collar, and let's work together on the widespread adoption of standards.

References

1. Fenner M, Crosas M, Grethe J, et al. A Data Citation Roadmap for Scholarly Data Repositories. Preprint bioRxiv 097196; https://doi.org/10.1101/097196 (2016).
2. McMurry J, Juty N, Blomberg N, et al. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. ***PLoS Biol.*** 15(6):e2001414 (2017).
3. Taylor CF, Field D, Sansone SA, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. ***Nat Biotechnol***. 26(8):889-96 (2008).
4. Empty rhetoric over data sharing slows science. Editorial. ***Nature*** 546, 327 (2017).

5. Sansone SA and Rocca-Serra P. Interoperability Standards - Digital Objects in Their Own Right. Wellcome Trust. Figshare https://doi.org/10.6084/m9.figshare.4055496.v1 (2016).
6. Field D, Sansone SA, Collis A et al. Megascience. 'Omics data sharing. ***Science***. 9;326(5950):234-6 (2009).
7. McQuilton P, Gonzalez-Beltran A, Rocca-Serra P et al. BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. ***Database*** (Oxford). 17;2016. pii: baw075. (2016).
8. McQuilton P, Gaudet P, Sansone SA. BioSharing survey. Figshare https://doi.org/10.6084/m9.figshare.3795810.v2 (2016).
9. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. ***Sci Data***. 15(3):160018 (2016).
10. Wilkinson MD, Sansone S-A, Schultes E, et al. A design framework and exemplar metrics for FAIRness. bioRxiv 225490; https://doi.org/10.1101/225490 (2017).

Acknowledgements

We thank all our collaborators and contributors to the current resource and its precursors (BioSharing and the Minimum Information about a Biomedical or Biological Investigation, MIBBI portal), in particular the chairs and the members of our international Advisory Board, the Force11 and RDA WGs and our Adopters community. We specifically thank key (past and present) contributors to the resource, notably Melanie Adekale, Delphine Dauga, Eamonn Maguire, Annapaola Santarsiero, and Chris Taylor. The authors are funded by grants awarded to S.A.-S. that include elements of FAIRsharing; specifically grants from the UK BBSRC and Research Councils (BB/L024101/1, BB/L005069/1), EU (H2020-EU.3.1, 634107, H2020-EU.1.4.1.3, 654241, H2020-EU.1.4.1.1, 676559), IMI (116060), and NIH (U54 AI117925, 1U24AI117966-01, 1OT3OD025459-01, 1OT3OD025467-01, 1OT3OD025462-01). S.A.-S. is funded also by the Oxford e-Research Centre of the University of Oxford.

Authors contributions

S-A.S. developed the concept and provided the strategic direction, and with P.R.-S. launched the initial portal, which was re-branded, curated, enriched and further developed by A.L., M.T., M.I. and A.G.-B. under the coordination of P.McQ. S-A.S. and P.McQ. led the working group under Force11 and the RDA. P.McQ. assembled the statistics, and S-A.S. and P.McQ wrote the manuscript with contributions from all authors.

Competing interested

The authors declare no competing financial interests.