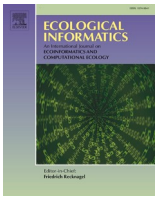
Ecological Informatics 64 (2021) 101372

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Integrating long-tail data: How far are we?

Kristin Vanderbilt [a,*], Corinna Gries [b]

[a] Environmental Data Initiative, University of New Mexico, Albuquerque, NM 87131, United States of America
[b] Environmental Data Initiative, University of Wisconsin, Madison, Madison, WI 53706, United States of America

ARTICLE INFO

Keywords:
Synthesis
Long-tail data
Data repositories
Semantics
Data harmonization

1. Background

Most ecological or environmental datasets, collected by individual researchers or small projects, are viewed as being from the "long tail of science" because they are low volume and formatted according to the needs of the individual scientist (Heidorn, 2008). These small, tabular, site-based datasets often document historical environmental conditions, providing a reference by which to assess environmental change. Through the integration of disparate datasets from many small projects, broader temporal and spatial scale research questions can be addressed, offering vast potential for new inferences about regional or global level processes. Integrating long-tail data can be daunting, however, because they may differ by temporal and spatial scales, units, semantics, data collection methodology, sampling design, and data organization. Hence, harmonizing heterogeneous datasets prior to analysis may be far more time-consuming than the data analysis itself (Anaconda, 2020; Wickham, 2014), and several mechanisms to streamline the process of integrating data have been proposed or implemented in the last decade.

Data repositories support and ease the process of data integration by providing FAIR (Findable, Accessible, Interoperable, and Reusable) data (Wilkinson et al., 2016). Before data can be integrated, they must be *findable*, and a repository usually offers several ways to query its corpus of metadata. Repositories that require extensive, standardized metadata and keywords from standard vocabularies may offer increased discoverability through faceted data searches (Diepenbroek et al., 2017). *Accessibility* of data is ensured when repositories 1) assign a persistent identifier such as a DOI to each dataset, 2) provide a clear data use license, and 3) provide manual or web services API data download capabilities. *Interoperability* is a function of the quality of metadata required by the repository. A repository that accepts detailed metadata in a standard format such as the Ecological Metadata Language (EML) (Fegraus et al., 2005), which describes the physical structure of a dataset, allows automated data extraction from the repository to port into workflows for data integration. The quality and richness of metadata have a direct bearing on data *reusability*. Metadata without elements such as parameters, units, methods and study design description may render a dataset of little use to a synthesis project. Lack of metadata is considered a risk factor for future data reusability (Mayernik et al., 2020).

One challenge to data integration is understanding the meaning of terms, or concepts, used in the metadata and data to describe parameter names, sampling and analytical methods. Within a given data integration project, a common set of terms and their definitions must be adopted that can be mapped to the source datasets to harmonize semantics across all datasets (Soranno et al., 2015). For datasets using an idiosyncratic set of terms, interpreting the meaning of parameter names may require time-consuming consultation with the data originator before datasets can be combined. Data integration would require less effort if data providers used standardized domain terminologies and definitions when selecting parameter names or assigning keywords. Although in some disciplines, such as organismal trait research, several controlled vocabularies are in use (e.g., Garnier et al., 2017; Walls et al., 2012), smaller projects that integrate trait data rarely refer to them (Schneider et al., 2019). In other disciplines, e.g., ecology, it may still be difficult to know which terminology is of highest quality or most appropriate to apply. Implementation of ontologies will further assist

* Corresponding author.
E-mail address: krvander@fiu.edu (K. Vanderbilt).

https://doi.org/10.1016/j.ecoinf.2021.101372
Received 17 July 2021; Accepted 19 July 2021
Available online 24 July 2021

data integration efforts by clarifying semantic relationships between concepts in the dataset, improving data discovery and interoperability (Kissling et al., 2018a, 2018b; Parr et al., 2016).

Data are most easily integrated and reused if they are provided in a community developed standard format with standard terminology. The Darwin Core Archive (DwC-A) is an example of a data standard that stores species data in a self-contained dataset along with a metadata file indicating how the data are organized (Wieczorek et al., 2012). Parameters in the data files are drawn from the Darwin Core glossary of terms, and the structured, XML-based metadata file in the DwC-A facilitates machine readability of the data. The Global Biodiversity Information Facility (GBIF) accepts data that are submitted as DwC-A files and then aligns them based on Darwin Core terms. GBIF can then enable integration of hundreds of millions of species occurrence records from many different sources through its data portal (Heberling et al., 2021).

Once data are prepared in such a community-accepted standard format, automation of data processing steps has the potential to greatly accelerate the pace of data integration. The lack of standardization in repository approaches to data discovery, filtering, retrieval, and processing, however, can make workflow implementation challenging, as illustrated by the following example from the biodiversity research community. Automated workflows are needed that can reliably reproduce the steps required to derive Essential Biodiversity Variables (EBV) (Pereira et al., 2013), which integrate large numbers of heterogeneous species occurrence records that vary across space, time, and taxa from many repositories. Hardisty et al. (2019) developed workflows to use GBIF and Atlas of Living Australia to find, filter, and then retrieve species occurrence data for further processing and merging into a common EBV data product. Although both repositories supply biodiversity data and offer similar features, each workflow required custom coding, expert advice, and the use of tools external to the repository to complete. The authors concluded that repositories need to cooperate to harmonize their infrastructure and integrate more tools to facilitate repeatable and efficient data processing.

A limiting factor for any data synthesis project is availability of relevant datasets. Many datasets from small-scale government agency or academic projects have never been contributed to repositories. There are many reasons why researchers do not share data (Kim and Zhang, 2015; Astell and Admin, 2018), among which is the lack of incentives to do so. This is changing as publishers and funders require data publication from authors and grant recipients, and because repositories now assign persistent identifiers to datasets, making it possible for dataset authors to receive credit for their dataset via a citation. Data citation offers other benefits, such as improved scientific reproducibility and provenance tracking. Yet datasets are frequently either not cited or cited incorrectly (Vannan et al., 2020). Increased awareness of the importance of this practice is needed so that data providers can get credit for contributing their data to a repository.

2. This special issue

In this special issue we share seven papers that describe current data synthesis efforts and consider mechanisms for accelerating the pace of data integration. The papers touch on the topics identified above, illustrating their efficacy, or suggesting needed changes to data management approaches by researchers or data repositories.

First, a project to harmonize quantitative individual plant-level trait data from multiple heterogeneous sources such as taxonomic revisions and ecological datasets is described (Lenters et al., 2021). The authors developed a workflow that accepts as inputs Excel files with non-standard formats and terminologies and a metadata form, filled out by the data provider. The metadata form facilitates linking of terms in the input dataset to standardized terms from ontologies such as the Plant Trait Ontology. R scripts identify errors in the input files, validate the metadata, integrate the input datasets, and generate four harmonized, machine-readable output files with semantic links to existing ontologies. The workflow was successfully used to integrate 15 palm datasets with nearly 140,000 individual plant trait measurements.

Second, Ely et al. (2021) describe development of a metadata and data reporting format for leaf-level gas exchange data, which are widely used in synthesis projects and for model parameterization. These data require specialized expertise to collect, are output in different formats by different instruments, and are rarely accompanied by metadata sufficient for unambiguous data interpretation. This project undertook to develop a unifying standard format for archiving these data to promote discoverability and reuse. By engaging over eighty data providers and data users, consensus was reached on a reporting format that includes guidance for variable names and definitions, file structure, units and metadata content. Archive of data consistent with this reporting format will improve useability, lessening the burden of data integration of leaf-level gas exchange data.

Third, Bond-Lamberty et al. (2021) describe a metadata and data reporting format for observations of soil-to-atmosphere carbon dioxide flux, or soil respiration. No centralized repository exists for these chamber-level measurements, and the aim of this project is to facilitate archiving the data in a machine actionable format that fosters data synthesis across multiple data sources. With community input, the authors selected standard variable names and definitions, and developed a suite of templates to help data providers format their data per the standard. The reporting standard is maintained in an open online github repository and will evolve as community needs change.

Fourth, O'Donnell et al. (2021) describe their process for harmonizing long-term greater sage-grouse monitoring data collected by state agencies in nine states in the western United States. The authors detail how they standardized the data across different sampling methodologies, terminologies, degrees of quality control and data file structures. An open source software tool was generated to automate the integration of the state datasets. The paper includes several figures and tables illustrating analyses performed with the synthesized dataset.

Fifth, Huber et al. (2021) explore how machine actionability of data and metadata varies across data repositories. Data synthesis would be more efficient if data and metadata could be read directly into analytical environments from persistent identifiers, without the need for intervention by a human. The authors surveyed multiple research infrastructures (RIs), such as PANGAEA and NEON, and found that dataset persistent identifiers resolve to human-accessible landing pages and do not provide machine-actionable links to data objects. They also observed that individual RIs develop specialized software libraries to allow connections between their APIs and the many programming languages used by scientists for data analysis. The landscape of tools used to access and ingest RI data and metadata into analytical platforms is thus heterogeneous and not standardized. The authors argue for future coordination among RIs to harmonize technologies and develop generic approaches to loading data into analysis tools, regardless of the programming language used. They offer a roadmap to reach this goal with specific implementation suggestions.

Sixth, O'Brien et al. (2021) discuss how raw ecological community data held in repositories can be processed by repository personnel into a harmonized format to facilitate data integration. Community observation data frequently differ with respect to data collection protocols and environmental sampling conditions, making them a challenge to synthesize. Data managers from the Environmental Data Initiative (EDI) repository and National Ecological Observatory Network (NEON) cooperated to develop a data model that accommodates different types of measurements, sampling designs, and taxonomic resolution. NEON and EDI have transformed 530 datasets into this format, called eco-comDP. Both infrastructures have developed workflows and an R code library (Smith and Sokol, 2021) that will regenerate the harmonized data products when the underlying raw data are updated. DwC-A files can also be created. An advantage of this approach to providing harmonized data is that it does not result in a separate database for the harmonized data, but rather includes the transformed datasets in the

repository that houses the raw data.

Finally, Agarwal et al. (2021) explore approaches for citing groups of datasets from the perspective of both data producers, who want credit for their data being used, and data users, who want a compact citation. Ensuring that all datasets used in a data integration effort are appropriately cited may be problematic if the number of dataset citations exceeds the space available for them in a paper. The authors discuss three current approaches—data collections, dynamic data citations, and data papers– for resolving citation and credit issues in the context of different use cases from their experiences with earth and environmental science data. They note that none of these methods solve all issues and propose that a 'container' with a unique identifier could contain citations for anything from dynamic data citations to data papers to supporting documentation about the resources in the 'container'.

3. Conclusions

This special issue illustrates that progress is being made toward making long-tail data less time-consuming and difficult to integrate. Based on these papers, we can offer some recommendations to scientists and data repositories whose goal is to make data more interoperable and reusable:

1) Use of domain-specific standard terminologies by both data providers and repositories can significantly lessen the pre-analysis time needed to harmonize datasets. When data providers do the work of mapping their data to standardized resources containing concepts and their definitions, they make the data easier to understand, support machine processing, and make the data more discoverable;
2) Employing a common data format for a given scientific domain, whether the formatting is done by the data provider or by the repository, facilitates machine actionability and reduces the need for tedious pre-harmonization dataset transformations;
3) Data repositories should coordinate technology development to provide a common set of tools to support data and metadata ingestion into arbitrary analytical platforms. This approach will allow repositories to benefit from economies of scale, and researchers will not have to develop different workflows for each data repository they use; and
4) For situations where datasets are too numerous to cite individually in a synthesis paper, the ecological informatics community needs to provide clear citation advice to ensure that data providers receive credit and data synthesizers have a succinct citation.

Declaration of Competing Interest

None.

Acknowledgements

Support for the authors and editors was provided by the US National Science Foundation to the Environmental Data Initiative (Grants #1931143 and #1931174) and to the Florida Coastal Everglades Long Term Ecological Research Program (Grant #DEB-9910514).

References

Agarwal, D.A., Damerow, J., Varadharajan, C., Christianson, D.S., Pastorello, G.Z., Cheah, Y.-W., Ramakrishnan, L., 2021. Balancing the needs of consumers and producers for scientific data collections. Ecol. Inform. 62, 101251. https://doi.org/10.1016/j.ecoinf.2021.101251.

Anaconda, 2020. State of Data Science 2020 [WWW Document]. Anaconda. https://www.anaconda.com/state-of-data-science-2020 (n.d., accessed 6.2.21).

Astell, M., Admin, S.N., 2018. Infographic - Practical Challenges for Researchers in Data Sharing. https://doi.org/10.6084/m9.figshare.5996786.v4.

Bond-Lamberty, B., Christianson, D.S., Crystal-Ornelas, R., Mathes, K., Pennington, S.C., 2021. A reporting format for field measurements of soil respiration. Ecol. Inform. 62, 101280. https://doi.org/10.1016/j.ecoinf.2021.101280.

Diepenbroek, M., Schindler, U., Huber, R., Pesant, S., Stocker, M., Felden, J., Buss, M., Weinrebe, M., 2017. Terminology supported archiving and publication of environmental science data in PANGAEA. J. Biotechnol. Bioinform. Solut. 261, 177–186. https://doi.org/10.1016/j.jbiotec.2017.07.016.

Ely, K.S., Rogers, A., Agarwal, D.A., Ainsworth, E.A., Albert, L.P., Ali, A., Anderson, J., Aspinwall, M.J., Bellasio, C., Bernacchi, C., Bonnage, S., Buckley, T.N., Bunce, J., Burnett, A.C., Busch, F.A., Cavanagh, A., Cernusak, L.A., Crystal-Ornelas, R., Damerow, J., Davidson, K.J., De Kauwe, M.G., Dietze, M.C., Domingues, T.F., Dusenge, M.E., Ellsworth, D.S., Evans, J.R., Gauthier, P.P.G., Gimenez, B.O., Gordon, E.P., Gough, C.M., Halbritter, A.H., Hanson, D.T., Heskel, M., Hogan, J.A., Hupp, J.R., Jardine, K., Kattge, J., Keenan, T., Kromdijk, J., Kumarathunge, D.P., Lamour, J., Leakey, A.D.B., LeBauer, D.S., Li, Q., Lundgren, M.R., McDowell, N., Meacham-Hensold, K., Medlyn, B.E., Moore, D.J.P., Negrón-Juárez, R., Niinemets, Ü., Osborne, C.P., Pivovaroff, A.L., Poorter, H., Reed, S.C., Ryu, Y., Sanz-Saez, A., Schmiege, S.C., Serbin, S.P., Sharkey, T.D., Slot, M., Smith, N.G., Sonawane, B.V., South, P.F., Souza, D.C., Stinziano, J.R., Stuart-Haëntjens, E., Taylor, S.H., Tejera, M.D., Uddling, J., Vandvik, V., Varadharajan, C., Walker, A.P., Walker, B.J., Warren, J.M., Way, D.A., Wolfe, B.T., Wu, J., Wullschleger, S.D., Xu, C., Yan, Z., Yang, D., 2021. A reporting format for leaf-level gas exchange data and metadata. Ecol. Inform. 61, 101232. https://doi.org/10.1016/j.ecoinf.2021.101232.

Fegraus, E.H., Andelman, S., Jones, M.B., Schildhauer, M., 2005. Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. Bull. Ecol. Soc. Am. 86, 158–168. https://doi.org/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2.

Garnier, E., Stahl, U., Laporte, M.-A., Kattge, J., Mougenot, I., Kühn, I., Laporte, B., Amiaud, B., Ahrestani, F.S., Bönisch, G., Bunker, D.E., Cornelissen, J.H.C., Díaz, S., Enquist, B.J., Gachet, S., Jaureguiberry, P., Kleyer, M., Lavorel, S., Maicher, L., Pérez-Harguindeguy, N., Poorter, H., Schildhauer, M., Shipley, B., Violle, C., Weiher, E., Wirth, C., Wright, I.J., Klotz, S., 2017. Towards a thesaurus of plant characteristics: an ecological contribution. J. Ecol. 105, 298–309. https://doi.org/10.1111/1365-2745.12698.

Hardisty, A.R., Belbin, L., Hobern, D., McGeoch, M.A., Pirzl, R., Williams, K.J., Kissling, W.D., 2019. Research infrastructure challenges in preparing essential biodiversity variables data products for alien invasive species. Environ. Res. Lett. 14, 025005 https://doi.org/10.1088/1748-9326/aaf5db.

Heberling, J.M., Miller, J.T., Noesgaard, D., Weingart, S.B., Schigel, D., 2021. Data integration enables global biodiversity synthesis. PNAS 118. https://doi.org/10.1073/pnas.2018093118.

Heidorn, P.B., 2008. Shedding light on the dark data in the long tail of science. Libr. Trends 57, 280–299. https://doi.org/10.1353/lib.0.0036.

Huber, R., D'Onofrio, C., Devaraju, A., Klump, J., Loescher, H.W., Kindermann, S., Guru, S., Grant, M., Morris, B., Wyborn, L., Evans, B., Goldfarb, D., Genazzio, M.A., Ren, X., Magagna, B., Thiemann, H., Stocker, M., 2021. Integrating data and analysis technologies within leading environmental research infrastructures: challenges and approaches. Ecol. Inform. 61, 101245. https://doi.org/10.1016/j.ecoinf.2021.101245.

Kim, Y., Zhang, P., 2015. Understanding data sharing behaviors of STEM researchers: The roles of attitudes, norms, and data repositories. Libr. Inf. Sci. Res. 37, 189–200. https://doi.org/10.1016/j.lisr.2015.04.006.

Kissling, W.D., Ahumada, J.A., Bowser, A., Fernandez, M., Fernández, N., García, E.A., Guralnick, R.P., Isaac, N.J.B., Kelling, S., Los, W., McRae, L., Mihoub, J.-B., Obst, M., Santamaria, M., Skidmore, A.K., Williams, K.J., Agosti, D., Amariles, D., Arvanitidis, C., Bastin, L., Leo, F.D., Egloff, W., Elith, J., Hobern, D., Martin, D., Pereira, H.M., Pesole, G., Peterseil, J., Saarenmaa, H., Schigel, D., Schmeller, D.S., Segata, N., Turak, E., Uhlir, P.F., Wee, B., Hardisty, A.R., 2018a. Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. Biol. Rev. 93, 600–625. https://doi.org/10.1111/brv.12359.

Kissling, W.D., Walls, R., Bowser, A., Jones, M.O., Kattge, J., Agosti, D., Amengual, J., Basset, A., van Bodegom, P.M., Cornelissen, J.H.C., Denny, E.G., Deudero, S., Egloff, W., Elmendorf, S.C., Alonso García, E., Jones, K.D., Jones, O.R., Lavorel, S., Lear, D., Navarro, L.M., Pawar, S., Pirzl, R., Rüger, N., Sal, S., Salguero-Gómez, R., Schigel, D., Schulz, K.-S., Skidmore, A., Guralnick, R.P., 2018b. Towards global data products of essential biodiversity variables on species traits. Nat. Ecol. Evol. 2, 1531–1540. https://doi.org/10.1038/s41559-018-0667-3.

Lenters, T.P., Henderson, A., Dracxler, C.M., Elias, G.A., Kamga, S.M., Couvreur, T.L.P., Kissling, W.D., 2021. Integration and harmonization of trait data from plant individuals across heterogeneous sources. Ecol. Inform. 62, 101206. https://doi.org/10.1016/j.ecoinf.2020.101206.

Mayernik, M.S., Breseman, K., Downs, R.R., Duerr, R., Garretson, A., Hou, C.-Y. Sophie, Committee, E.D.G.I. (EDGI) and E.S.I.P. (ESIP) D.S, 2020. Risk assessment for scientific data. Data Sci. J. 19, 10. https://doi.org/10.5334/dsj-2020-010.

O'Brien, M., Smith, C.A., Sokol, E.R., Gries, C., Lany, N., Record, S., Castorani, M.C.N., 2021. ecocomDP: A flexible data design pattern for ecological community survey data. Ecol. Inform. 64, 101374 https://doi.org/10.1016/j.ecoinf.2021.101374.

O'Donnell, M.S., Edmunds, D.R., Aldridge, C.L., Heinrichs, J.A., Monroe, A.P., Coates, P. S., Prochazka, B.G., Hanser, S.E., Wiechman, L.A., Christiansen, T.J., Cook, A.A., Espinosa, S.P., Foster, L.J., Griffin, K.A., Kolar, J.L., Miller, K.S., Moser, A.M., Remington, T.E., Runia, T.J., Schreiber, L.A., Schroeder, M.A., Stiver, S.J., Whitford, N.I., Wightman, C.S., 2021. Synthesizing and analyzing long-term monitoring data: a greater sage-grouse case study. Ecol. Inform. 63, 101327. https://doi.org/10.1016/j.ecoinf.2021.101327.

Parr, C.S., Schulz, K.S., Hammock, J., Wilson, N., Leary, P., Rice, J., Corrigan Jr., R.J., 2016. TraitBank: practical semantics for organism attribute data. Semant. Web 7, 577–588. https://doi.org/10.3233/SW-150190.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C.,

Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M., Wegmann, M., 2013. Essential biodiversity variables. Science 339, 277–278. https://doi.org/10.1126/science.1229931.

Schneider, F.D., Fichtmueller, D., Gossner, M.M., Güntsch, A., Jochum, M., König-Ries, B., Provost, G.L., Manning, P., Ostrowski, A., Penone, C., Simons, N.K., 2019. Towards an ecological trait-data standard. Methods Ecol. Evol. 10, 2006–2019. https://doi.org/10.1111/2041-210X.13288.

Smith, Colin, Sokol, Eric, 2021. ecocomDP: Work with Datasets in the Ecological Community Design Pattern. R package version 1.0.0. https://CRAN.R-project.org/package=ecocomDP.

Soranno, P.A., Bissell, E.G., Cheruvelil, K.S., Christel, S.T., Collins, S.M., Fergus, C.E., Filstrup, C.T., Lapierre, J.-F., Lottig, N.R., Oliver, S.K., Scott, C.E., Smith, N.J., Stopyak, S., Yuan, S., Bremigan, M.T., Downing, J.A., Gries, C., Henry, E.N., Skaff, N. K., Stanley, E.H., Stow, C.A., Tan, P.-N., Wagner, T., Webster, K.E., 2015. Building a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse. GigaScience 4. https://doi.org/10.1186/s13742-015-0067-4.

Vannan, S., Downs, R., Meier, W., Wilson, B.E., Gerasimov, I., 2020. Data Sets Are Foundational to Research. Why Don't We Cite Them? [WWW Document]. Eos. https://eos.org/opinions/data-sets-are-foundational-to-research-why-dont-we-cite-them (accessed 6.3.21).

Walls, R.L., Athreya, B., Cooper, L., Elser, J., Gandolfo, M.A., Jaiswal, P., Mungall, C.J., Preece, J., Rensing, S., Smith, B., Stevenson, D.W., 2012. Ontologies as integrative tools for plant science. Am. J. Bot. 99, 1263–1275. https://doi.org/10.3732/ajb.1200222.

Wickham, H., 2014. Tidy data. J. Stat. Softw. 59, 1–23. https://doi.org/10.18637/jss.v059.i10.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D., 2012. Darwin core: an evolving community-developed biodiversity data standard. PLoS One 7, e29715. https://doi.org/10.1371/journal.pone.0029715.

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18.