A Systematic Review of Data Management Platforms

Michael Boch[1], Stefan Gindl[1], Alan Barnett[2], George Margetis[3], Victor Mireles[4], Emmanouil Adamakis[3], and Petr Knoth[5]

[1] Research Studios Austria Forschungsgesellschaft mbH Salzburg, Österreich
{michael.boch,stefan.gindl}@researchstudio.at
[2] Dell Technologies, County Cork, Ireland
Alan.Barnett@Dell.com
[3] Foundation for Research and Technology - Hellas, Greece
{gmarget,madamakis}@ics.forth.gr
[4] Semantic Web Company, Vienna, Austria
victor.mireles-chavez@semantic-web.com
[5] Knowledge Media institute, The Open University, UK
petr.knoth@open.ac.uk

Abstract. This paper systematically reviews a set of well-established data management platforms and compares their functionality. We derived an initial criteria catalogue from existing research work and extended it based on the input gathered through several expert interviews. Finally, we applied this criteria catalogue to a set of data management platforms. The contribution of this work is (i) an up-to-date criteria catalogue to systematically assess the feature-richness of data management platforms, generalizable to related use-cases (e.g. data markets), and (ii) the systematic review of a selected set of data management platforms along these criteria. This work lays the foundation for future research in this area, being subject to periodic re-evaluation to also include developments and improvements of the platforms.

Keywords: Data management platform · Data management system · Data portal.

1 Introduction

Data management platforms (DMPs) are an essential tool for the management, storage, maintenance, and distribution of data assets, in both academic, research and industrial [11] contexts [1]. Low-level approaches to data management, e.g. directly using the file system of any computer, are common for small amounts of data, but quickly show their limitations when significant amounts of data are stored, processed, and provided to users. In such systems, it can be hard to handle data redundancy, keep data assets up to date, make them FAIR [24], and provide a central point for their processing. The use of specialized DMPs proves exceptionally useful, as they allow to make data accessible to a larger audience, provide functionality for searching, curating, preserving and matching data assets, and integrate with both computational and organizational processes. Their

functionalities include storing, cataloguing, searching, and ideally also matchmaking between data assets, (optionally) assignment of unique persistent identifiers, and archiving. DMPs often also feature full-text search engines indexing both metadata and content, as well as workflows for curating or reviewing deposited data assets. They are crucial in an era where research and industrial applications benefit from diverse data, coming from different domains, backgrounds, and sources. DMPs allow organizations to create a central location for data assets. Finally, DMPs provide functionalities that are the basis of data marketplaces, commercial ventures for the exchange of diverse data assets.

In this work, we survey, explore, and compare the functionalities of widely used DMPs and answer the following **research questions**: (i) "What are suitable criteria for comparing data management platforms?", and (ii) "What are the differences and similarities between data management platforms?"

1.1 Related Work

The language used to classify computing infrastructure for data management is diverse and inconsistent. One general category deals with the management of research data (e.g. [1]), considering software for the depositing, preservation, curation, and publication of research data. This category can include both document repositories re-purposed for data (e.g. DSpace[6]) as well as open-data publishing systems (e.g. CKAN[7]). Target users of such systems are primarily researchers. An interesting development in this area is the abandonment of the purely archivist perspective on research data management, in favour of a more functional one which includes data re-usability by humans and machines[22]. Research data management has become one of the key focus areas of the Open Science movement[20, 8], as embodied in the FAIR principles.

A seemingly orthogonal approach to data management, but which nonetheless shares many of the functional and architectural requirements, is that of commercial or industrial data management. In this setting, organizations are interested in aggregating data from internal databases, management systems (e.g. CRM), and external data brokers in order to advance the digitization of their operations and power business intelligence applications. These systems also require the construction of data catalogues which power search and discovery, as well the standardization of descriptions, access methods and provenance keeping[5].

Industrial DMPs can also serve as a stepping stone for the exchange of data between organizations for commercial purposes. While there is no general definition of the term *data marketplace* [21], its commercial and business dimensions must be supported by the management of the data assets. The term "data management platform" has also been used in the field of digital marketing (e.g. [12]), with a decidedly different meaning to the one we deal with here. In the following, we give a differentiation of terms commonly used in this domain:

- **Open Data Portal:** is a web catalogue for freely available and reusable data (open data) with the goal to facilitate the discoverability of digital assets.

[6] www.dspace.com, accessed Nov 10, 2021
[7] ckan.org, accessed Nov 10, 2021

- **Data Marketplace:** a location to buy and sell data online[16]. Data marketplaces act as an intermediary offering an infrastructure between the provider and the buyer of the data [15].
- **Data catalogue/data dictionary:** is a repository of metadata containing metadata for all application data of a company and the relationships between the different data objects.
- **Data Warehouses:** used since the 1980s [4], data warehouses refer to a central data storage to bring together data from different databases. The goal is fast retrieval of data that is understandable for the business units of companies [13]. Data warehouses use predefined schemas [6] to increase performance in data retrieval and high data quality.
- **Data lake:** store data in its raw format without loss of information, but no means of quality assessment of the data. The concept is called "schema-on-read" since the data is loaded unstructured and transferred into a schema when retrieved [6]. This allows to store large amounts of data at low cost, but slows down data access [7]. Special data formats such as Apache Parquet allow an increase in reading speeds[8]. The unstructured type of data requires versatile storage technologies.

We see DMPs as software (online or offline) for the archiving, lookup, and exchange of data assets, not confined to research data (in contrast to [2]). Data assets refer to datasets, but also resources such as services (e.g. for analysis or visualization of data) or machine learning models.

https://www.overleaf.com/project/6160370edfdc58b28ed72b1a

2 Methodology

The methodology of this research is based on the design science research process according to [19] (see Fig. 1).

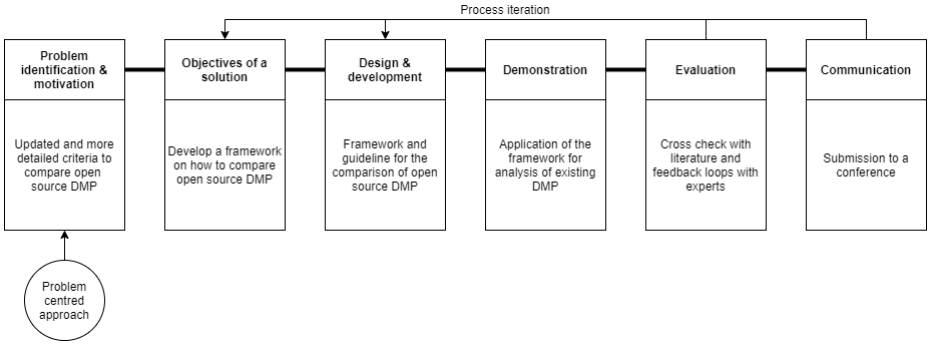

Fig. 1. The applied design science research process derived from [19].

[8] https://parquet.apache.org/

The goal is to develop a comparative criteria catalogue for open-source DMPs from an initial set of criteria. The criteria catalogue was created via a concept-centric literature review adapted from [23] and [14], to include concepts from [17]. The resulting set of criteria leverages existing research and extends it with new criteria. This resulted in the demonstration phase of the DSR process creation of a theoretical criteria catalogue, ready to assess DMPs and compare them. The catalogue was later evaluated via a cross-check with literature and feedback loops with experts.

The overall evaluation phase uses a methodological triangulation, leveraging the results of (i) existing previous work, (ii) expert assessments, as well as (iii) practical experience. The process consisted of five consecutive steps, as described in the following section and depicted in Fig. 1.

2.1 Criteria in Detail

This section discusses the individual criteria in detail. The criteria developed by [1] and [2] served as the starting point for the criteria catalogue. To achieve a higher level of detail for technical aspects some criteria were split up. The literature review served as the starting point for the additional criteria. Features were also derived from concepts presented in the papers. The criteria identified were then evaluated and validated in several expert rounds. Figure 2 gives an overview of this procedure.

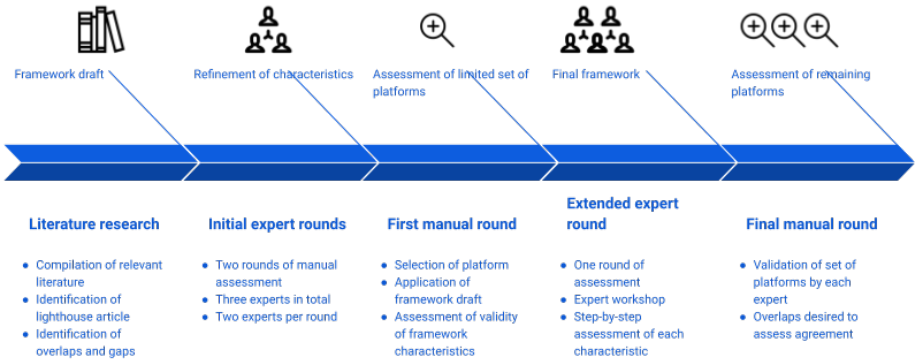

Fig. 2. Methodology for the creation of the criteria catalogue.

The evaluation of most criteria was adapted from [17] (see Table 1). Their open-source nature allows for inclusion of missing features. However, this can be associated with a considerable amount of additional effort. The mentioned experts are the authors of the paper.

1. **Literature research:** The first step consisted of the examination of existing literature. The identification of a lighthouse article [2] provided the starting point which served as a basis for extension in the later phases. The outcome of this step was the definition of a preliminary criteria catalogue.

2. **Initial expert round:** Three experts validated the preliminary criteria catalogue in a first round. They discussed the suggested dimensions and criteria and agreed on their validity and applicability for the use case of DMPs, resulting in an extension as well as a reduction/consolidation of the criteria.
3. **First manual round:** A single annotator, selected from the group of experts, applied the criteria catalogue to a limited set of DMPs. This helped to pre-test and gain practical experience with the criteria. The result of this step was a reviewed criteria catalogue, with comments and suggestions based on the gained practical experience.
4. **Extended expert round:** Five experts discussed the outcome of the previous steps. The outcomes of this manual round provided valuable input and practical examples for making a final decision on the criteria to be included in the final criteria catalogue.
5. **Final manual round:** The experts applied the final criteria catalogue to the identified set of DMPs, iterating over the criteria and evaluating them for each DMP. Each expert assessed one or more DMPs and the sets of DMPs per expert were overlapping. The experts used an adapted desk research method for this step. This comprised the consultation of existing online documentation of the respective platforms, available on the DMPs' websites and online code repositories. The result of this step was the completed categorization of DMPs along the dimensions of the criteria catalogue.

Table 1. Scoring System adapted from [17].

Name	Description
Y (Yes)	Assigned if natively supported.
∼	Assigned if the criterion was partially fulfilled.
P (Plugin)	Assigned if available as a plugin.
N (No)	Assigned if the criterion was not fulfilled.
–	Assigned if the criterion was not applicable.
?	Assigned if it was not possible to assess the mentioned criterion.

3 Evaluation

In the previous sections, criteria were collected and successfully applied. Based on [19] this section explains the evaluation phase based on a cross-check with literature and feedback loops with experts. Altogether there were two feedback loops with six experts. In each round, the criteria were discussed in detail and, if necessary, adjusted based on the experts' opinions. Between the feedback loops, the criteria were applied to the identified DMPs of the shortlist. Each DMP was evaluated by at least 2 experts. The cross-check of results was also part of the discussions in the feedback loops.

In the initial expert round, a set of criteria derived from the work of Amorim et al. [1, 2] were used as the foundation of the criteria catalogue. Their criterion "API" was expanded within the initial expert round. Concepts based on scientific literature served as a starting point. "RESTful access service" derived from the functional requirements of [10] can also be assigned to the point API and

Table 2. Assessed criteria catalogue (as of October 2021).

Feature	**CKAN**	**Dataverse**	**DSpace**	**ePrints**	**Fedora**	**InvenioRDM**	**Omeka**
Version	2.9	5.6	7	3.4	6	6.0	3.1
Installation wizard	N	Y	N	N	Y	N	N
Installable from repository	Y	Y	Y	Y	Y	Y	Y
Distributed as Container Image	Y	Y	N	Y	Y	Y	Y
Licence	AGPL	Apache	BSD	LGPL	Apache	MIT	GPL
Ecosystem of extensions	Y	Y	N	Y	N	N	Y
Internationalization support	Y	Y	Y	P	N	N	Y
Multi factor authentication	Y	N	Y	N	Y	N	N
Authorization (access control)	Y	Y	Y	Y	Y	Y	Y
Federated identity/Single sign-on	P	Y	Y	Y	Y	Y	N
Review system	N	N	N	N	N	N	N
Maturity	Y	Y	∼	Y	Y	Y	Y
Programming language	Python	Java	Java, Angular	Perl, Javascript	Java	Python	PHP
Billing/payment system	N	N	N	N	N	N	P
Documentation	Y	Y	Y	Y	Y	Y	Y
Cloud storage interface	P	Y	Y	Y	Y	Y	N
Exporting schemas	N	Y	Y	Y	Y	Y	N
Dataset versioning	P	Y	N	Y	Y	Y	Y
Dataset license specification	Y	Y	Y	N	N	Y	Y
Billing schema per dataset	N	N	N	N	N	N	N
Faceted search	Y	Y	Y	Y	Y	Y	Y
Duplicate detection	N	N	N	P	N	N	N
Established ontologies	P	Y	Y	Y	Y	Y	Y
Customized ontologies	∼	N	Y	N	N	N	P
Domain-specific vocabularies	Y	P	Y	N	Y	Y	Y
Customized vocabularies	∼	P	Y	N	N	Y	Y
Data provenance management	P	N	Y	N	N	N	N
Diverse data types	Y	Y	Y	Y	Y	Y	Y
Exporting and RDF serialisation	Y	P	N	Y	Y	N	Y
Data services	N	N	N	N	Y	Y	N
Unstructured data	Y	Y	Y	Y	Y	Y	Y
Harvesting / Crawling of datasets	Y	Y	Y	Y	N	N	P
Batch access API	P	Y	Y	N	Y	N	N
Transactional access API	Y	Y	Y	Y	Y	Y	Y
Analytical access API	P	P	Y	N	Y	Y	N
Data visualisation	Y	Y	N	N	N	N	N
SWORD compliance	N	Y	Y	Y	Y	N	N
Data anonymization	N	N	N	N	N	N	N

therefore confirms the relevance of the criterion. Different types of access were also addressed in [9]. Based on the FAIR principles from [17] access to metadata via HTTP(s) was generalised to "Analytical access API". All criteria adapted from [17] are derived on the FAIR principles discussed in the paper such as "Data provenance management" and "Exporting and serialisation". The latter can also be seen as a generalisation of "Exporting schemas" by [1]. The criterion "Batch access API", is based on the criterion "OAI-compliance" of [1]. Since the focus of the paper has moved away from research to a more broad perspective the criteria "No unique identifier", "Pre-reserving DOI" and "Embargo period" of [1] was removed. However, these criteria should be addressed in future work with a focus on research data management. The criterion "Registered repositories" of [2] collected the number of instances according to OpenDOAR. It was removed since only a part of the platforms is listed and thus no general comparison with all platforms was possible. "Complex installation or setup" was removed in the initial expert round because it was unclear how to assess it. The question of when an installation is complex could not be answered clearly. Based on this, the criteria "Installation wizard", "Installable from a repository" and "Container" were added during the initial expert round. Since the paper only reviews self-hosted platforms, the criterion "Maintenance costs" was removed. Due to the focus on platforms the criteria "Deployment" and "Storage location" of [1] were replaced with "Cloud storage interface". Since all the analysed DMPs are platforms and open-source and therefore customizable, the criteria "Customization" was removed. Instead, the criterion "Ecosystem of Extensions" was introduced. "Multi factor authentication" and "Authorization" are derived from [17] and from the functional requirements of [3]. "Schema flexibility" of [1] was split up into "Established ontologies" and "Domain-specific vocabularies" based on the concepts discussed in [17] and [10] . "Versioning" of [1] was renamed to "Dataset versioning" and "Record licence specification" of [2] to "Dataset license specification" for better comprehensibility. "Faceted Search" was derived from [17]. "Diverse data types" was derived from [18] and conceptually from [10]. The criterion "Licence", "Maturity", "Programming language", "Billing/payment system", "Documentation", "Duplicate detection", "Data services", "Unstructured data", "Harvesting /Crawling of datasets", "Data visualisation", "SWORD compliance" were added based on expert opinions and to then be evaluated in the extended expert round. The following adjustments were applied during the extended expert round. The majority of the criteria from the initial expert round were confirmed. The criteria "Established ontologies" and "Domain-specific vocabularies" were expanded ("Customized ontologies" and "Customized vocabularies"). "Validation", as suggested by [1], the option for researchers to assert correct metadata structure was removed since "Review system" conceptually derived from [9] was seen as a more general version of it. Derived from the FAIR principles [17], interoperability with data preservation systems was also considered. This was also excluded due to the shift of focus compared to [1]. These criteria should still be addressed future works addressing research data management. One aspect from the initial expert round was an in-

terface to LDAP. However, this was discarded as it is more relevant for internal solutions within companies. [9] focuses on data marketplaces, but some criteria are also applicable for DMPs. One criterion suggested by [9] that was included in the first draft of the criteria catalogue during literature search was "Encryption" and the question of whether and what kind of encryption exists. This point was discarded in the extended expert round as it was too unspecific for the present study according to the experts. However, security is an important topic and is therefore reflected in other criteria. The criteria "Federated identity" and "Single sign-on" were combined into one criterion and "Maturity" is a refinement of concepts from the initial expert round.

4 Conclusion & Future Outlook

The goal of this research was to establish a criteria catalogue for comparing DMPs, ultimately containing 38 criteria. In order to show its applicability, we applied it to DMPs using a design science research process. We considered a total of 18 DMPs, and examined seven of them using the criteria catalogue. We excluded non open-source platforms in this work. Additionally, Magda.io[9] was excluded from the study, because it is still in an early stage. However, it already supports several data types, data visualisation, federated authentication, duplicate detection, authorization features, and an API. Future work should also include Magda.io. Greenstone[10] and Open Data Catalog[11] were excluded because of the inactivity of their public repository. Amorim et al. excluded Fedora[12] in [2] stating that its focus lies more on being a fully customizable framework than being a finished out-of-the-box solution which makes it not comparable with other DMPs. We included Fedora because it is suitable for a detailed analysis based on our exclusion and inclusion criteria, even when only considering the core services. In addition, our criteria catalogue with the shift in focus also allows a more detailed mapping of the differences.

All platforms are theoretically extensible, but not all have an ecosystem of extensions. DSpace and InvenioRDM do not offer a hub for extensions yet. However, since InvenioRDM has only recently been completed, this may yet emerge. For ePrints, extensions can be found online, but due its small number, this cannot yet be classified as an ecosystem. Compared to the other DMPs, Omeka[13] (variant Omeka S was examined) has the most specialised focus. It was designed for exhibition of cultural heritage objects, even though its focus has now shifted.

None of the examined DMPs is out-of-the-box suitable as a data market, since no platform offers a billing/payment system natively and only Omeka offers this with a plugin. A billing system at dataset level is not available on any platform. Each of the examined DMPs also offers an API. However, not all platforms support all types of APIs. All DMPs considered offers the necessary

[9] magda.io, accessed Nov 10, 2021
[10] www.greenstone.org, accessed Nov 10, 2021
[11] github.com/azavea/Open-Data-Catalog, accessed Nov 10, 2021
[12] duraspace.org/fedora, accessed Nov 10, 2021
[13] omeka.org, accessed Nov 10, 2021

basic functionalities. Data can be archived and is searchable, as all plattforms also have advanced search functionalities.

This work created and applied a criteria catalogue to assess features of DMPs. We developed the criteria catalogue in several feedback loops using design science research. The suitability was demonstrated with the application of the criteria onto selected DMPs. It was successfully applied to 7 different DMPs. The comparison reviewed the differences and similarities of the DMPs and it pointed out the high quality of all the DMPs reviewed and that the differences depend on the respective main field of application.

Future work will focus on the application of the criteria catalogue to new DMPs and update it to include the characteristics of new software releases. This ensures that future software trends will get addressed. The year 2021 in particular saw an increase in new developments. InvenioRDM, a data repository offshoot from Invenio, had its first production ready release. DSpace released version 7 which brought a new Angular-based UI and extended REST API. And for ePrints, originally planned as a publication repository, the entire code base was renewed with version 3.4 and released in different versions including an Open Data portal variant. The catalogue can is usable as a blueprint for a comparison of service-based DMPs. Furthermore, future work will apply the criteria catalog to compare DMPs exclusively for the research and science domain.

Acknowledgement

This work was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 871481

References

1. Aguiar, J., da Silva João Rocha, Carvalho, R.C.A.R., Castro: A comparative study of platforms for research data management: Interoperability, metadata capabilities and integration potential. pp. 101–111. Springer International Publishing (2015)
2. Amorim, R.C., Castro, J.A., da Silva, J.R., Ribeiro, C.: A comparison of research data management platforms: architecture, flexible metadata and interoperability. Universal Access in the Information Society **16**, 851–862 (2017)
3. Brandao, A., Mamede, H.S., Goncalves, R.: Trusted data's marketplace. pp. 515–527. Springer International Publishing (2019)
4. Devlin, B.A., Cote, L.D.: Data warehouse: From architecture to implementation (1996)
5. Dibowski, H., Schmid, S., Svetashova, Y., Henson, C., Tran, T.: Using semantic technologies to manage a data lake: Data catalog, provenance and access control (11 2020)
6. Dorschel, J.: Praxishandbuch Big Data. Springer Gabler, Wiesbaden, 1 edn. (2015)
7. Fang, H.: Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. pp. 820–824 (2015)
8. Fenner, M., Crosas, M., Grethe, J.S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M., Clark, T.: A data citation roadmap for scholarly data repositories. Scientific Data **6**, 28 (2019)
9. Fruhwirth, M., Rachinger, M., Prlja, E.: Discovering business models of data marketplaces. pp. 5736–5747 (1 2020), 53rd Hawaii International Conference on System Sciences, HICSS 2020 ; Conference date: 07-01-2020 Through 10-01-2020

10. Gregory, A., Chen, C., Wu, R., Miller, S., Ahmad, S., Anderson, J.W., Barrett, H., Benedict, K., Cadol, D., Dascalu, S.M., Others: Efficient model-data integration for flexible modeling, parameter analysis and visualization, and data management. Frontiers in Water **2**, 2 (2020)
11. Hameri, A.P., Nihtilä, J.: Product data management—exploratory study on state-of-the-art in one-of-a-kind industry. Computers in Industry **35**, 195–206 (1998)
12. Hüllmann, J.A., Sivakumar, A., Krebber, S.: Data management platforms: An empirical taxonomy (2021)
13. Kimball, R., Ross, M.: The Data Warehouse Toolkit. Wiley, 3 edn. (2013)
14. Kitchenham, B.A.: Procedures for performing systematic reviews (2004)
15. Koutroumpis, P., Leiponen, A., Thomas, L.D.W.: The (unfulfilled) potential of data marketplaces (2017)
16. Lange, J., Stahl, F., Vossen, G.: Datenmarktplätze in verschiedenen forschungsdisziplinen: Eine Übersicht. Informatik-Spektrum **41**, 170–180 (2018)
17. Langer, A., Bilz, E., Gaedke, M.: Analysis of current rdm applications for the interdisciplinary publication of research data. vol. 2447. CEUR-WS.org (2019)
18. Nayek, J.K.: Evaluation of open data government sites: A comparative study. Library Philosophy and Practice p. 1 (2018)
19. Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A design science research methodology for information systems research. Journal of Management Information Systems **24**, 45–77 (2007)
20. Pontika, N., Knoth, P., Cancellieri, M., Pearce, S.: Fostering open science to research using a taxonomy and an elearning portal. Association for Computing Machinery (2015)
21. Stahl, F., Schomm, F., Vossen, G., Vomfell, L.: A classification framework for data marketplaces. Vietnam Journal of Computer Science **3**, 137–143 (2016)
22. Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R.R., Duerr, R., Haak, L.L., Haendel, M., Herman, I., Hodson, S., Hourcle, J., Kratz, J.E., Lin, J., Nielsen, L.H., Nurnberger, A., Proell, S., Rauber, A., Sacchi, S., Smith, A., Taylor, M., Clark, T.: Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science (2015)
23. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly **26**, xiii–xxiii (2002)
24. 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., 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.: The fair guiding principles for scientific data management and stewardship. Scientific Data **3**, 160018 (2016)