IJDC | *Conference Pre-print*

Finding a Repository with the Help of Machine-Actionable DMPs: Opportunities and Challenges

Simon Oblasser
Vienna University of technology

Thomas Miksa
Vienna University of Technology

Asanobu Kitamoto
National Institute of Informatics

Abstract

Finding a suitable repository to deposit research data is a difficult task for researchers since the landscape consists of thousands of repositories and automated tool support is limited. Machine-actionable DMPs can improve the situation since they contain relevant context information in a structured and machine-friendly way and therefore enable automated support in repository recommendation.

This work describes the current practice of repository selection and the available support today. We outline the opportunities and challenges of using machine-actionable DMPs to improve repository recommendation. By linking the use case of repository recommendation to the ten principles for machine-actionable DMPs, we show how this vision can be realized. A filterable and searchable repository registry that provides rich metadata for each indexed repository record is a key element in the architecture described. At the example of repository registries we show that by mapping machine-actionable DMP content and data policy elements to their filter criteria and querying their APIs a ranked list of repositories can be suggested.

Submitted 13 December 2019 ~ Accepted 19 February 2020

Correspondence should be addressed to Tomasz Miksa, Favoritenstraße 16, 1040 Wien. tmiksa@sba-research.org

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Repositories play a key role in sharing and long-term preservation of research data and are therefore crucial for the promotion of Open Science. Many funders and publishers mandate or recommend to deposit research data underlying a publication in a suitable repository. The landscape of research data repositories, consisting of several thousand institutional, disciplinary or general repositories, is hard to comprehend and researchers have difficulties in selecting the 'right one'. In a recent survey from Springer Nature (Stuart et al., 2018) researchers were asked about their problems in data sharing and 33% stated they did not know which repository to use.

The selection of a repository can depend on many factors such as the established community standards, type of data, metadata or licensing. Funders and publishers might impose additional requirements for a repository to be recommended such as providing open access, supporting the FAIR principles (Wilkinson et al., 2016) or being trustworthy (Science Europe, 2018). The FAIR data expert group, advising the European Commission, recommends repositories to seek CoreTrustSeal[1] (CTS) certification (Hodson et al., 2018). To deposit data in a CTS certified repository is currently not required, but is recommended by many funding bodies.

To keep track of all the repository options out there, repository registries like DataCite's re3data[2] or Jisc's OpenDOAR[3] maintain curated lists of repositories which can be searched, filtered and programmatically accessed. Rich metadata associated with each repository allow to draw repository profiles and make them comparable (Kindling et al., 2017).

A common question in DMPs is about data sharing and which repositories will be used to deposit the datasets being created during the research project. Since DMPs are written in an early phase of the project, often before data is being created, this might be the first time researchers have to answer this question. Hence, support in repository selection at this stage is very useful for researchers and also helps to identify conditions for data deposit early on. Context information provided in a DMP such as type, format and size of data, chosen metadata standards, researcher affiliations or the funder's data policy can be used to guide and automate repository recommendation and enable more accurate results.

In this paper, we present how repository recommendation can be automated with the use of machine-actionable DMPs, in order to reduce the effort for researchers of finding the suitable repository. We describe which components are needed to implement this vision. We analyse to what extent the already existing components, such as registries or policies, support the proposed architecture and which new developments and community efforts are needed. To do so, we identify fields in maDMPs relevant for repository selection and map them to the existing repository registries. We also analyse existing funder requirements, which can be linked from maDMPs, to identify rules influencing repository selection. Based on that analysis, we discuss which elements necessary to implement automated repository recommendation are in place and which still need development and further community discussion.

This paper is structured as follows. Section 2 presentes related work. Section 3 describes the proposed solution. Section 4 describes the mapping of maDMP fields to filters in repository registries. It also presents how funder policies can be expressed in rules and how these map to repository selection filters. Section 5 presents discussion. Conclusion and future work appear in the last section.

[1] https://www.coretrustseal.org/

[2] https://www.re3data.org/

[3] http://v2.sherpa.ac.uk/opendoar/

Related work

Common practice and tools for repository recommendation

Various stakeholders such as funders, publishers or journals are providing lists of recommended repositories, e.g. listed in FAIRsharing[4] (Husen et al., 2017). Predominantly publishers like the Public Library of Science[5] (PLOS) or Springer Nature[6] provide curated lists of recommended repositories for data deposits linked to journal publications on their websites. In both examples about 100 repositories grouped by discipline are recommended. Repositories play a key role in sharing and long-term preservation of research data and are therefore crucial for the promotion of Open Science. Many funders and publishers mandate or recommend to deposit research data underlying a publication in a suitable repository.

Recent projects like the DARIAH Data Deposit Recommendation Service[7] (DDRS) or the DataCite Repository Finder[8] are built on top of re3data by querying its Elasticsearch API. DDRS as part of an open humanities data platform, guides through a set of questions (country, humanities discipline) in order to filter for repositories that accept data deposit and assign persistent identifiers (PIDs) (Buddenbohm, 2017). Inspired by DDRS, Repository Finder as part of the Enabling FAIR Data initiative[9], provides a similar service in the domain of earth, space and environmental sciences (Witt, 2019). As a motivation for their service they modelled the complexity of selecting a data repository as a decision tree (Enabling FAIR Data Community, 2018). Repository Finder filters re3data for disciplinary, open access repositories that accept data deposit and issue PIDs. As current limitations of the tool they identified the lack of information about the user's affiliation, so no relevant institutional repositories could be suggested. They also identified limitations related to the metadata[10] of each repository record in re3data such as the completeness of information, e.g. to check FAIR practices or the granularity of information, e.g. search repositories by type of data.

Another tool to find a suitable repository in the domain of biological sciences is the data submission wizard[11] provided by the European Bioinformatics Institute (EMBL-EBI). In contrast to the previously mentioned tools, the EMBL-EBI tool maintains its own list of 20 data repositories. The wizard asks questions about the type of data, access control, used instruments etc. to make a repository recommendation.

Machine-actionable DMPs and repository recommendation

The idea of repository recommendation within data management planning has already been discussed. Witt and Giarlo (2012) described the vision of an automated repository recommendation by using contextual information and keywords from the DMP.

Machine-actionable DMPs (maDMPs), which capture information in a structured way by using machine-readable data formats, standards and controlled vocabulary instead of free-form text, are an ideal source of contextual information and provide a new way to automate repository recommendation.

Repository recommendation was described as a major use case in the white paper about maDMPs (Simms et al., 2017) and recommending repositories or tools was rated as high priority functionality for a DMP tool in a survey about the Horizon 2020 DMP template (Grootveld et al., 2018).

[4] https://fairsharing.org/recommendations/
[5] https://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories
[6] https://www.springernature.com/gp/authors/research-data-policy/recommended-repositories
[7] https://ddrs-dev.dariah.eu/ddrs/
[8] https://repositoryfinder.datacite.org/
[9] http://www.copdess.org/enabling-fair-data-project/
[10] http://doi.org/10.2312/re3.008
[11] https://www.ebi.ac.uk/submission/

Automating repository recommendation using maDMPs

Machine-actionable DMPs offer the opportunity to capture and bundle relevant information from multiple sources, such as: funder, researcher, repository, to make repository recommendation more accurate. This information can be reused to facilitate repository selection by automatically setting specific search criteria.

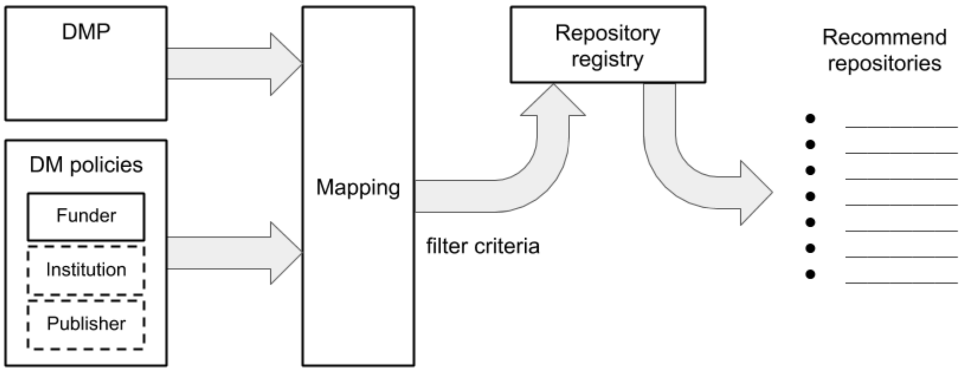

Figure 1. Information coming from a DMP in a structured way, such as type, format and size of data, selected metadata standards, researcher affiliation, licensing etc. combined with data policy information, e.g. funder requires the repository to issue DOIs and recommends CoreTrustSeal certification can be mapped to filter criteria and used to query a repository registry for matching repository entries.

Figure 1 depicts the architecture supporting automated filtering of repositories based on information from maDMPs. DMP information, such as type, format and size of data, selected metadata standards, researcher affiliation or licensing can be mapped to filter criteria of a repository registry. The same applies to data policies, e.g. a funder requires that the repository issues DOIs and recommends a CoreTrustSeal certification, which can be mapped to filter criteria of a repository registry. The mapping from the machine-actionable attributes to the repository registry's filter criteria needs to be specified. In some cases it can be a 1:1 mapping, this means the DMP uses the same controlled vocabulary as the repository registry does for its filter criteria, e.g. metadata standards specified by the DCC schema. In other cases, mapping heuristics may be applied, e.g. data types in a DMP may be described by their mime-type, while a repository registry uses another vocabulary to describe a filter for content types. Once the filter criteria are determined the repository registry's API can be queried for matching repositories.

The described architecture is independent of any specific service. However, a prerequisite for a repository registry is that it provides programmatic access and allows to filter and search for repository entries based on their properties.

In the remainder of this paper we analyse and discuss what is needed to realize the architecture presented in this section.

Mapping maDMPs, policies, and repository registries

The common standard for maDMPs (Miksa et al., 2019) is based on stakeholder consultations and represents information needed and provided by all stakeholders involved in the process of data management. In this paper, we analysed each field in terms of relevance for the repository selection. Then for each field identified as relevant we looked for equivalent filters in re3data and OpenDOAR registries. The goal was to identify to what extent the information contained in the model can be used to automatically filter relevant repositories.

Direct mapping of maDMP fields to registry filters

Table 1 shows the results of the mapping exercise. We have identified that values for some fields overlap, for example, Distribution/data_access set to restricted has an exact same value at re3data as in the maDMP. In such cases the information from the maDMP can be used directly to set filter criteria in the registry.

There are also fields where there is a possible match, for example Dataset/type, because the fields exist in the maDMP and in at least one of the registries. However, the maDMP specification allows any value here, re3data has its own vocabulary, while OpenDOAR another one. Hence, there is a need for common vocabulary of content types to be used across the systems, otherwise the filtering may not be efficient.

There are also fields that exist in the maDMP, but do not have any counterparts in the registries, e.g format or byte_size. Although restricting format types supported may be one of the preservation requirements, none of the registries phrases it explicitly. Similar for the size limit of accepted submissions. This information can mostly be found in human-readable versions of repository policies, but is currently not available in any of the registries. Another reason may be lack of cost models explicitly used by repositories.

For fields where no mapping was found, we additionally checked FAIRsharing for missing values. Only in the case of Metadata/metadata_id we identified a mapping. Using metadata_id that references a specific metadata standard, we can find repositories (FAIRsharing: databases) that follow this standard.

Table 1. Mapping of maDMP fields to re3data and OpenDOAR filters.

maDMP field	re3data	OpenDOAR	Comment
DMP/ ehtical_issues_exist Dataset/ personal_data Dataset/ sensitive_data	databaseAccess	policies/ metadata_policy policies/ data_policy	Partial alignment possible - repositories with restricted access or with policies that control access to specific items
Dataset/ type	contentType	content_types	Common vocabulary needed
Dataset/	keywords	---	Only high level match possible

maDMP field	re3data	OpenDOAR	Comment
keyword			- keywords describing dataset must match keywords describing repository
Distribution/ format	(contentType)	(content_types)	No direct mapping, heuristic approximation can be used in specific cases

(Table 1 continued)

Distribution/ byte_size	---	---	No corresponding filter found
Distribution/ data_access	dataAccess	policies/ data_policy	Mapping possible, maDMPs use almost the same controlled vocabulary as re3data
License/ license_ref	dataLicense	(policies/ data_policy)	Partial alignment possible - re3data defines it as license of data existing in the repository - not of what licenses can be assigned. OpenDOAR models policies and does not refer to specific ones, e.g. CC-BY
Cost/*	restrictionType / {registrationFeeRequired}	---	re3data can filter repositories where registration fee is necessary. Otherwise no information on costs.
Metadata/ metadata_id	metadataStandards	---	Partial alignment - re3data indicates what metadata is used by the repository. This does not preclude uploading different metadata as part of the submission.

Indirect information provided by maDMPs

In the second stage we focused on information linked by a maDMP that can be obtained automatically, but is not modelled directly as a field within the maDMP. We identified the following fields as relevant entry points that allow us importing further requirements for repository selection:

- Funding/funder_id
- Funding/grant_id

- Contributor/*

Based on the funder and grant ids we can identify funder's policies applicable. Furthermore, based on the roles of contributors, e.g. researchers working at a specific university, we can identify applicable institutional policies.

Most of the European research funders follow the Science Europe (2018) recommendations on data management and repository selection. We analysed Science Europe recommendations and also an open data policy of one of the Austrian funders[12] to identify specific requirements that can be used in automation of repository selection. We identified the following requirements:

1. Repository must assign PIDs
2. Repository must assign licenses to individual items
3. License should be open
4. Access type must be specified
5. Repository must be listed in re3data
6. Repository must be institutional, discipline-specific or interdisciplinary
7. Repository should be CTS certified

Currently, there are no machine-actionable policies in place. To make policies machine-actionable, we can express their relevant parts as concrete rules. For example, a high level requirement to allow for persistent identification of data, can be expressed as a rule stating that repository must use DOI, arks, or handles. We follow here the LEARN project[13] methodology that distinguishes between policies, taboos, principles and rules.

It is possible to define rules for all identified requirements, because they are either binary options, e.g. license is open or not, or there is a controlled vocabulary of options that can be checked, e.g. classification of repositories.

Table 2 shows the mapping of the policy derived rules to filters in the re3data and OpenDOAR registries. For all of them we were able to find a matching filter. This can indicate that the funder requirements are closely related to the functionality offered by the registries nowadays. The rules derived from funder policies overlap with information already included in maDMPs, with just a few exceptions, such as recommendation to use CoreTrustSeal certified repositories.

An alternative approach to the one described in this section is implemented by FAIRsharing that also aims at connecting funder or publisher requirements to specific databases. It does not model specific policy content but simply creates lists of repositories compliant with a specific funder or publisher policy. This saves the effort of breaking down policies into specific rules. However, it is based on manual annotation of repositories and requires manual and periodical review of both policies and new repositories.

Table 2. Mapping of research funder policy derived rules to registry filters.

Rule derived from policy	re3data	OpenDOAR	Comment
Repository must assign PIDs	pidSystems	---	Filtering possible

[12] https://www.fwf.ac.at/en/research-funding/open-access-policy/open-access-to-research-data/

[13] http://learn-rdm.eu/en/about/

Rule derived from policy	re3data	OpenDOAR	Comment
Repository must assign licenses to individual items	dataLicense	policies/ data_policy	Filtering possible

(Table 2 continued)

License should be open	dataLicense	policies/ data_policy	Filtering possible. re3data: classification of existing licenses into open/closed licenses is needed. OpenDOAR: does not specify exact policy, e.g. CC-BY, but describes what is allowed to be done with data. uses also “access_phrases” to state e.g. “free_open_access”
Access type must be specified	dataAccess	policies/ data_policy	Filtering possible
Repository must be listed in re3data	yes	--	Filtering possible
Repository must be institutional, discipline-specific or interdisciplinary	types	repository_type	Filtering possible
Repository should be CTS certified	certificates	---	Filtering possible

Discussion

The analysis shows that maDMPs contain information that is relevant in view of repository recommendation. Thus, by reusing information from maDMPs, it is possible to reduce the amount of manual input from researchers and other stakeholders involved in research data management.

However, the analysis also showed that despite many community efforts there are still open gaps that need to be closed before a fully operational system can be deployed. To systematise the discussion, we present our findings by iterating over the principles for machine-actionable DMPs (Miksa et al., 2019) that we found relevant for automated repository recommendation.

#3 Make policies (also) for machines, not just for people

Existing funder policies are still meant for people only. They have relatively few requirements influencing selection of a repository that can be broken down into specific rules. These can be manually created when deploying automated recommendation services within a well-defined context, e.g. a university where most of the funding comes from a limited number of funders. FAIRsharing has recently proposed a set of criteria (McQuilton et al., 2019) used to describe funder or publishers policies.

#4 Describe - for both machines and humans - the components of the DM ecosystem

This principle can be established by repository registries like re3data or OpenDOAR which provide descriptions of repository records in machine-readable formats (JSON, XML), using controlled vocabulary defined in schemes. For example, re3data provides a XML schema[14] with controlled vocabulary describing the filter criteria, which helps in establishing the mapping from machine-actionable attributes. Both registries provide APIs allowing machines to access their contents.

However, the problem is in alignment of ways information is represented. For example, OpenDOAR expresses policies in a machine-actionable way, by defining actions that may apply to data, e.g. data can be reused by anyone. Re3data on the other hand provides a list of licenses that a given repository uses. Thus, it does not focus on the details of what specific licenses allows for, like OpenDOAR does, but uses references to common license types, e.g. CC-BY. This example shows that both registries provide information on reuse, but in a different way.

Mappings between specific services and maDMP fields are needed. No global or universal mapping is possible at the moment. This limits the interoperability and creates dependence on a specific registry. A common way of describing information for machines and humans on topics like access condition, licenses or costs is needed.

#5 Use PIDs and controlled vocabulary

PIDs and controlled vocabularies are necessary to refer to specific entities or resources. Examples of PIDs and controlled vocabulary in the context of repository recommendation are ORCID[15] (researcher), ROR[16] (affiliation), Crossref funder registry[17] (funder), PUID of PRONOM[18] registry (data description e.g. automated by FITS[19]), DCC metadata standards[20], re3data DOI (repository).

Using PIDs for funders can help in identifying relevant policies. Referencing metadata standards by PIDs facilitates finding relevant repositories listed by FAIRsharing.

There is still a need for common vocabularies. For example, each registry uses its own vocabulary for content type and these are not the only vocabularies that exist in this scope, e.g. COAR vocabulary of resources[21]. A common vocabulary used by maDMPs, repository registries and licenses can improve finding relevant repositories.

#6 Follow a common data model for maDMPs

The common standard for machine-actionable DMPs already exists (Miksa et al., 2019). A standardized maDMP created by different tools can be processed by services, such as a repository recommender.

---

14 http://doi.org/10.2312/re3.008

15 https://orcid.org/

16 https://ror.org/

17 https://www.crossref.org/services/funder-registry/

18 http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

19 https://projects.iq.harvard.edu/fits

20 http://rd-alliance.github.io/metadata-directory/

21 http://vocabularies.coar-repositories.org/documentation/resource_types/

Future revisions of the standard should add more controlled vocabularies, e.g. content type vocabulary, as discussed above.

The common standard contains already more information than registries can use for mapping. For example, maDMPs can contain information on costs or size of data. This information cannot be currently used when searching for repositories.

Further challenges

In repository recommendation there are different stakeholders with diverse interests. Repository operators want their repository to get ranked high in order to increase the likelihood of submissions and the collection of associated fees and increase their market share. Funders and publishers want high impact of their funded and published work and therefore prefer broadly supported and recognized repositories with high visibility and reusability of the research data while maintaining long-term accessibility. Researchers want to find a repository supporting their data in the best way while keeping efforts minimal and maximize visibility of their research and increase citation. A repository recommendation system wants to deliver good results for a variety of needs, DMPs and policies while providing a maximum of transparency, non-bias and be inclusive for new repositories. A successful repository recommendation system will have to take all stakeholder interests into consideration.

Another challenge is the ranking of repositories fulfilling the same criteria. How to order the result list? Which repository should come first? Very likely the first repository will be used by default in most cases and thus can receive most of submissions. How does this affect the existence of other repositories? These and other similar questions must be answered when implementing automated repository recommendation system.

From a technical point of view any ranking scheme can be implemented. As a search engine, re3data uses Elasticsearch, which weighs and sorts the search results by relevance to the search query. The way ranking is implemented in re3data is purely based on technical aspects using Lucene's Practical Scoring Function[22]. Re3data allows to combine filter options with full-text search. Therefore, the relevance score is calculated by the combination of different methods such as the Boolean model, Term Frequency (TF)/Inverse Document Frequency (IDF), and the vector space model.

However, with Elasticsearch it is possible to control the scoring by applying custom scoring functions. The scoring functions can be applied to filtered subsets of the search results and boost results matching specific filters with adjustable weights. In other words, this allows to apply any kind of weighting schemes for filter criteria. For instance, a funder specifies that it is more important that a repository has CTS certification than it is supporting versioning, this can be expressed with different weights for each of the criteria. In this manner it is possible to gradually differentiate between must-have and nice-to-have criteria of a repository and reflect this in the ranking of search results. There might be different needs by different users, also depending on the field of study. Custom ranking schemes could help meet custom requirements by dynamically adjusting the scoring and influence the ranking of relevant repositories to increase the accuracy of recommendation results.

In this paper we propose a content-based recommendation system where we seek to find a best match between maDMPs, policies and repositories indexed in repository registries. However, it is possible to implement other kinds of recommendation systems. For instance, we could take the repository selections of other users into account and make recommendations based on similarity to other users. This collaborative-filtering approach requires a substantial amount of users before useful recommendations can be made. Also the introduction of new repositories is problematic since it suffers from a cold start problem where user data is missing in the beginning. Many successful recommender systems use a combination of different methods to achieve a high accuracy though.

[22] https://www.elastic.co/guide/en/elasticsearch/guide/current/practical-scoring-function.html

Conclusion

This paper dealt with the problem of automated repository selection for research data. It proposed an architecture in which information from machine-actionable DMPs is reused to automatically filter repository registries to find data repositories where research data described by the maDMP can be deposited. The goal was to reduce workload imposed on researchers by narrowing down the selection to repositories that support well data produced by researchers and meet funders' criteria.

We analysed to what extent the already existing components, such as repository registries or policies, support the proposed architecture and which new developments are needed.

The analysis showed that maDMPs contain information that is relevant in view of repository recommendation.

The existing funders' policies are not machine-actionable yet. They can be broken down into machine-actionable rules though. Thus, the rules can be mapped to the existing repository registry filters to facilitate search for relevant repositories.

The existing repository registries support access to information by machines. Unfortunately, there are various standards and vocabularies for describing properties of repositories. For example, different vocabularies are used to describe supported content types or different approaches are used to inform about the terms of access and licensing.

This analysis showed that the proposed architecture is feasible and can be implemented. However, further standardisation and service improvements are needed to fully utilize the information provided by maDMPs and to increase the quality of recommendations by being able to utilize information from different registries and policies that are expressed in a compatible way.

The common standard for maDMPs have other potential applications. The automated repository recommendation described here is one of them. The future work will also focus on other applications of the standard, such as integration with data repositories to automate data ingest.

Acknowledgements

This work was generously supported by the National Institute of Informatics (NII) International Internship Program. This research was carried out in the context of the Austrian COMET K1 program and publicly funded by the Austrian Research Promotion Agency (FFG) and the Vienna Business Agency (WAW).

References

Buddenbohm, S., De Jong, M., Minel, J., Priddy, M., Larousse, N., & Moranville, Y. (2017). D 7.2 Design and Sustainability Plan for an Open Humanities Data Platform: Concept for a Data Deposit Recommendation Service. DARIAH. halshs-01531337v2

Enabling FAIR Data Community, Duerr, R., Kinkade, D., Witt, M., & Yarmey, L. (2018). Data Repository Selection Decision Tree for Researchers in the Earth, Space, and Environmental Sciences. Zenodo. doi:10.5281/zenodo.1475430

Grootveld, M., Leenarts, E., Jones, S., Hermans, E., & Fankhauser, E. (2018). OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans. Zenodo. doi :10.5281/zenodo.1120245

Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., … Mietchen, D. (2018). Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data. doi:10.2777/1524

Husen, S. E., de Wilde, Z. G., de Waard, A., & Cousijn, H. (2017). Recommended versus Certified Repositories: Mind the Gap. Data Science Journal, 16, 42. doi:10.5334/dsj-2017-042

Kindling, M., Pampel, H., van de Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., Witt, M., Schirmbacher, P., Bertelmann, R., Scholze, F. (2017). The Landscape of Research Data Repositories in 2015: A re3data Analysis, D-Lib, 23, 3/4 doi:10.1045/march2017-kindling

McQuilton, P., Sansone, S.-A., Cousijn, H., Cannon, M., Chan, W. M., Carnevale, I., … Threlfall, J. (2019, October 28). FAIRsharing Collaboration with DataCite and Publishers: Data Repository Selection, Criteria That Matter. doi:10.17605/OSF.IO/N9QJ7

Miksa, T., Simms, S., Mietchen, D., & Jones, S. (2019). Ten principles for machine-actionable data management plans. PLOS Computational Biology 15(3): e1006750.doi:10.1371/journal.pcbi.1006750

Miksa, T., Walk, P., & Neish, P. (2019). RDA DMP Common Standard for Machine-actionable Data Management Plans. doi:10.15497/rda00039

Science Europe. (2018). *Practical Guide to the International Alignment of Research Data Management*. Retrieved from https://www.scienceeurope.org/media/jezkhnoo/se_rdm_practical_guide_final.pdf

Simms, S., Jones, S., Mietchen, D., & Miksa, T. (2017). Machine-actionable data management plans (maDMPs). Research Ideas and Outcomes. doi:10.3897/rio.3.e13086

Stuart, D., Baynes, G., Hrynaszkiewicz, I., Allin, K., Penny, D., Lucraft, M., & Astell, M. (2018). Whitepaper: Practical challenges for researchers in data sharing (Version 1). figshare. doi:10.6084/m9.figshare.5975011.v1

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data3:160018 doi:10.1038/sdata.2016.18

Witt, M., & Giarlo, M. J. (2012). *Databib: IMLS LG-46-11-0091-11 Final Report (White Paper)*. Libraries Reports. Paper 2. Available online: https://docs.lib.purdue.edu/libreports/2

Witt, M. et al. (2019). Connecting Researchers to Data Repositories in the Earth, Space, and Environmental Sciences. In: Manghi P., Candela L., Silvello G. (eds) Digital Libraries: Supporting Open Science. IRCDL 2019. Communications in Computer and Information Science, vol 988. Springer, Cham. doi:10.1007/978-3-030-11226-4_7