



# Towards a Toolbox for Automated Assessment of Machine-Actionable Data Management Plans

COLLECTION:  
DATA MANAGEMENT  
PLANNING ACROSS  
DISCIPLINES AND  
INFRASTRUCTURES

RESEARCH PAPER

]u[ubiquity press

TOMASZ MIKSA

MAREK SUCHÁNEK

JAN SLIFKA

VOJTECH KNAISL

FAJAR J. EKAPUTRA

FILIP KOVACEVIC

ANNISA MAULIDA NINGTYAS

ALAA EL-EBSHIHY

ROBERT PERGL

\*Author affiliations can be found in the back matter of this article

## ABSTRACT

Most research funders require Data Management Plans (DMPs). The review process can be time consuming, since reviewers read text documents submitted by researchers and provide their feedback. Moreover, it requires specific expert knowledge in data stewardship, which is scarce. Machine-actionable Data Management Plans (maDMPs) and semantic technologies increase the potential for automatic assessment of information contained in DMPs. However, the level of automation and new possibilities are still not well-explored and leveraged. This paper discusses methods for the automation of DMP assessment. It goes beyond generating human-readable reports. It explores how the information contained in maDMPs can be used to provide automated pre-assessment or to fetch further information, allowing reviewers to better judge the content. We map the identified methods to various reviewer goals.

## CORRESPONDING AUTHOR:

**Tomasz Miksa**

SBA Research, AT; TU Wien, AT  
[miksa@ifs.tuwien.ac.at](mailto:miksa@ifs.tuwien.ac.at)

## KEYWORDS:

maDMPs; automation;  
evaluation; funder; FAIR; RDM

## TO CITE THIS ARTICLE:

Miksa, T, Suchánek, M, Slifka J, Knaisl V, Ekaputra FJ, Kovacevic F, Ningtyas AM, El-Ebshihy, A and Pergl, R. 2023. Towards a Toolbox for Automated Assessment of Machine-Actionable Data Management Plans. *Data Science Journal*, 22: 28, pp. 1–13. DOI: <https://doi.org/10.5334/dsj-2023-028>

A data management plan (DMP) describes how research data is created, managed, shared, and preserved. Most funding bodies worldwide require DMPs for research projects (Miksa, Oblasser & Rauber 2021). The DMPs are reviewed by reviewers appointed by funders. Institutions, such as universities, also offer reviewing DMPs by research support staff, for example, before the DMP is submitted to funders.

The reviews are in most cases done manually, that is, reviewers read text documents submitted by researchers and provide their feedback. There are some standardized checklists for reviewers, for instance, Science Europe provides its evaluation rubric.<sup>1</sup> Given the scope of DMPs and the heterogeneity of practices regarding research data management across disciplines, it is hard for reviewers to be experts on all aspects, such as metadata standards, repositories, licensing, etc. Thus, the quality of feedback mostly depends on the reviewer's expertise and may lack required insight, overview, and objectivity.

Machine-actionable DMPs (maDMPs) (Miksa et al. 2020; Miksa et al. 2021) provide a structured way of organizing information contained in traditional DMPs. For instance, they make it explicit what datasets are created, where they will be published, what metadata will be used, and under which license. There is already a preliminary work done that shows how relevant information can be filtered and presented to reviewers by using SPARQL queries (Foidl et al. 2021). However, the potential for automatically validating information contained in DMPs is still not understood and leveraged by the broader research data management (RDM) community. For example, there are no systems in place that use maDMPs to facilitate the work of reviewers.

This paper aims to provide a toolbox of automation approaches for DMP assessment. We go beyond generating human-readable reports from machine-actionable DMPs. We investigate methods that help in assessing the quality of information provided, such as the extent to which decisions described by a DMP lead to FAIR data (Wilkinson et al. 2016), or whether the requirements by a specific funder are met. In our investigations, we consider the possibilities provided by semantic web technologies, as well as the tools and services that facilitate research data management. We map the identified methods to reviewers' goals to identify what kind of checks they support and what still remains the reviewers' task.

The paper is structured as follows. Section 2 describes the settings in which we consider the proposed methods, and presents the breakdown of reviewers' goals. Section 3 describes some selected methods for the automated assessment of DMPs. Section 4 maps the identified methods to the goals of reviewers. In Section 5, we discuss their limitations and give an outlook on further future research direction. Section 6 presents related work. We conclude the paper in Section 7.

## REQUIREMENTS AND USAGE SCENARIOS

In this section, we describe the current practice of reviewing the DMPs, and outline two possible scenarios that we consider feasible solutions that use maDMPs to improve the process. All other permutations can be derived from these scenarios. For simplicity, we focus on these two. We also identify the key goals of a DMP review. We will use these goals to drive our research on methods that facilitate the work of reviewers, and to evaluate how the identified methods contribute to these goals.

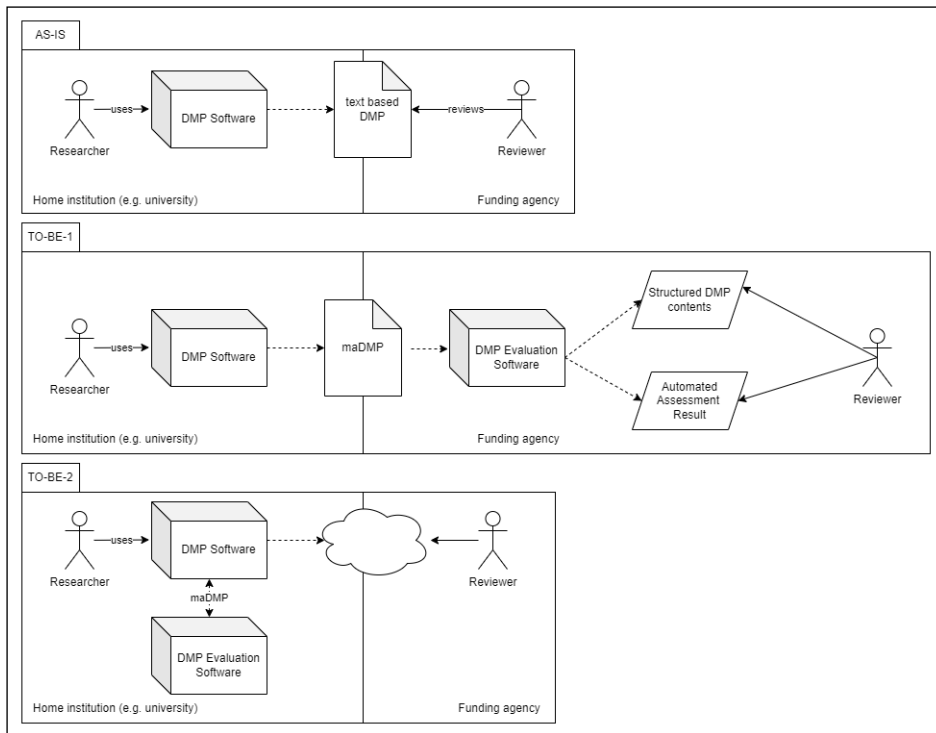
### USAGE SCENARIOS

The possibilities and benefits of machine-actionability have not been understood well by the RDM community, yet. This leads to new solutions trying to fine-tune the existing inefficient solutions, instead of re-thinking the bigger picture to improve the DMP review process.

Figure 1 depicts three different scenarios for DMP reviewing. The first one, labelled as AS-IS, depicts the typical setting currently in use by most funders. A researcher uses a DMP Software, such as DMP Tool, DAMAP, Argos, DS-Wizard, etc., to generate a PDF containing answers to

---

<sup>1</sup> [https://www.scienceurope.org/media/4brkxe5/se\\_rdm\\_practical\\_guide\\_extended\\_final.pdf](https://www.scienceurope.org/media/4brkxe5/se_rdm_practical_guide_extended_final.pdf).



**Figure 1** Scenarios depicting current situation (AS-IS) and two possible scenarios for automation of maDMP evaluation (TO-BE).

questions defined in a funder’s template. Reviewers evaluate the DMP by reading the answers from the text document.

The scenarios labelled as *TO-BE-1* and *TO-BE-2* depict potential settings that we consider in our work. In *TO-BE-1*, the researcher still uses the DMP Software to produce a DMP, but, compared to the current practice, the software sends the maDMP to the funder. Funders have a dedicated software for evaluation of maDMP. The software provides structured, human-readable information and metrics to pre-assess the answers in the maDMP. For instance, it displays the level of FAIRness of datasets described by the maDMP, so that the reviewers do not have to check that manually. The reviewer still makes the final judgment, but based on more solid information. This example shows that machine-actionability not only helps in exchanging and structuring information, but that it also enables machines to take actions based on this information, such as calculating the FAIRness level.

The scenario labelled as *TO-BE-2* is similar to the *TO-BE-1* scenario. The main difference is that the DMP evaluation software is used at the researcher’s side to provide feedback on the quality of the DMP, before it is sent to the funder, either as a traditional DMP or as a maDMP. In case the funders do not want to adopt any changes on their side, then the researcher’s side can still benefit from the automated checks. The symbol of the cloud used in the *TO-BE-2* scenario means that what is transferred to the funder is undefined, i.e. whether it is a traditional DMP or a maDMP. This has no impact on this scenario and depends on the specific setting, for example, how the stakeholders want to exchange information. In case the funders do not want to adopt any changes on their side, then the researcher’s side can still benefit from the automated checks.

## DMP REVIEW GOALS

The process of writing and reviewing a DMP does not have to be considered only from the perspective of fulfilling funder requirements, since DMPs can also be used or mandated in other settings. For example, writing a DMP can be an internal requirement of universities for their students, or companies can use DMPs as their internal documents to have a better overview of data used in their projects. Despite the fact that the majority of existing resources and use cases focus on checking whether DMPs fill funders’ requirements, we believe that checking compliance with funders’ requirements is not the only goal of the reviewers. The Science Europe Evaluation Rubric states that the work of the reviewer is to ‘assess whether the information provided in the DMP is sufficient to ensure that the research team will manage data as expected’ (Europe

2021). Based on this formulation, our own experience and discussions with fellow reviewers, we broke down this high-level objective into more specific goals:

- **G1. Completeness:** the coverage of all relevant aspects of research data management in the DMP. It can happen that DMPs miss some sections, e.g. no information on ethical aspects is provided. In other cases, DMP can provide only partial information, e.g. licenses are defined for only a subset of datasets listed in the DMP, and for unknown reasons the rest is undefined.
- **G2. Feasibility:** the possibility to put all DMP content into practice through concrete actions. This can include the identification of inconsistencies with a DMP. For example, a DMP does not plan sufficient storage for the data that will be produced.
- **G3. Quality of actions:** the relevance and effectiveness of the actions listed in the DMP and performed according to it. In other words, to assess whether what was described can be or was (depending on the phase) implemented according to community standards. For example, whether the data is FAIR, can be openly accessed, etc.
- **G4. Non-ambiguity:** clear and non-ambiguous formulation of the DMP. This is especially relevant for the non-machine-actionable parts of DMPs, that is, the parts containing text with verbal explanation, e.g. motivating the use of specific tools, techniques, services, etc.

In the remainder of the paper, we focus on methods dealing with machine-actionable information; language check of nonstructured text is out of scope for our investigations. Hence, methods described in Section 3 address goals G1-G3.

## METHODS FOR AUTOMATED ASSESSMENT OF DMPs

The paper does not suggest one specific system, tool, or solution to be used. Instead, it discusses a range of approaches that we found relevant and useful in the context of DMP assessment. It should facilitate and encourage the development of tools for automated DMP assessment.

DMPs are ‘living documents’, which means that towards the beginning of the project they describe planned actions, while towards the end they focus on describing the actions taken. The methods discussed below can be applied in different stages of the DMP lifecycle.

maDMPs heavily rely on information stored in other systems, such as grant databases, repositories, curated registries, and research knowledge graphs, among others. They use URLs and persistent identifiers to refer to this information (Miksa et al. 2019). Exploring these references is crucial in achieving real automation and machine-actionability. Most of the methods described in this section focus on exploiting them.

Figure 2 presents an overview of methods explored by us in order to create a toolbox for automated assessment of maDMPs.

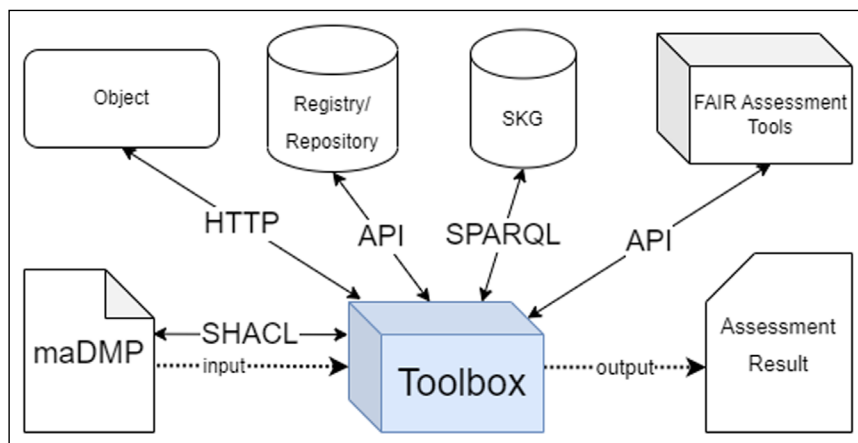


Figure 2 Overview of automation methods for DMPs.

## RDF-BASED VALIDATION

The emergence of the RDF serialization of maDMPs (Cardoso et al. 2022) opens up the possibility of utilizing semantic web-based validation of maDMPs. In their paper, Cardoso et al. provide an example maDMP validation using Shape Expression (ShEx) (Boneva et al. 2017), demonstrating

the validation capabilities of the semantic web technologies. Recently, W3C has approved recommendation of the Shapes Constraint Language (SHACL)<sup>2</sup> for RDF data validation, which provides a W3C standard alternative for ShEx.

These RDF-based validation methods facilitate maDMP content validation based on defined constraints. SHACL, in particular, allows users to describe constraints as shapes which contain descriptions of their targets—the nodes that they intend to validate. A target could be all instances of a particular RDF class, subjects or objects of a particular RDF property, or an explicit list of nodes. SHACL supports definitions of various constraint types. We provide the following examples of constraints that are particularly relevant to the maDMPs:

- **Class or data types:** to check if a property value is compatible with the allowed value types, e.g., ‘the cost value should be of type integer or float’.
- **Cardinality of property values:** to check whether the number of occurrences for a certain maDMP property is contained between the minimum and the maximum number of values allowed, e.g., ‘the number of title of a DMP should be exactly one’.
- **String regular expression matching:** to validate a property value against a given regular expression, e.g., ‘The URL of the system hosting should follow the following regex pattern: `^\\w+: (\\/) {0,2} [^\\s]+$`’.
- **Controlled vocabularies for property values:** to validate a property value against a closed list of allowed values, e.g., ‘the dataset distribution access value should be one of the following string: ‘open’, ‘shared’ or ‘closed’.

```
@prefix madmp: <https://w3id.org/madmp/terms#> .
@prefix ex: <https://w3id.org/dcso/id/example/> .
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

# ...
[] sh:property [
  sh:path madmp:contact_id;
  sh:name "The Contact ID Schema";
  sh:minCount 1; sh:maxCount 1;
  sh:property [
    sh:path madmp:identifier_type;
    sh:name "The DMP Contact Identifier Type Schema";
    sh:description "ID type. Allowed values: orcid, isni, openid, other";
    sh:minCount 1; sh:maxCount 1;
    sh:in ("orcid"^^xsd:string "isni"^^xsd:string
          "openid"^^xsd:string "other"^^xsd:string);
  ];
].
```

**Listing 1** An SHACL shape excerpt for validating DMP Contact Identifier Type (i.e., one of ‘orcid’, ‘isni’, ‘openid’ or ‘other’).

To demonstrate such functionality, we have developed the maDMP-Ontology Toolkit (MadPot)<sup>3</sup> with two main functionalities: (i) to allow transformation between maDMP JSON serialization to RDF and vice-versa, and (ii) to allow validation of maDMP ontology instances based on the specification provided in the DMP-Common-Standard specification. An excerpt of the SHACL validation shape is shown in Listing 1.

We evaluate MadPot on the examples maDMP JSON files provided from the DMP-Common-Standard GitHub page,<sup>4</sup> and the result shows that MadPot can detect issues within maDMP files successfully previously mentioned, i.e., *class/datatype*, *cardinality*, *regular expressions*, and *controlled vocabulary* validations. Further, the use of RDF and SHACL would further custom validation rules to be enforced for certain maDMPs, e.g., for specific institutions of funding agencies.

2 <https://www.w3.org/TR/shacl/>.

3 <https://github.com/fekaputra/MadPot>.

4 <https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard>.

One of the simplest approaches that does not require complex processing of information is identifier analysis. maDMPs contain identifiers that are mostly resolvable links, e.g. URIs that overlap with URLs, or DOIs that are in form of a URL. There are two types of checks for the links used in maDMPs that can help in the assessment:

- **Existence of linked objects:** to check whether the provided link actually points to an existing resource or whether it is broken. For example, this can be used to check whether a DOI for a dataset exists, i.e. whether the dataset was deposited into a repository.
- **Link whitelist/blacklist:** to check whether the DMP links to an allowed/forbidden resource.

To implement an evaluator tool for link existence, it is enough to check whether the HTTP response is in the successful range (2xx) or in the error range (4xx, 5xx). If the HTTP response is 3xx, the evaluator should first follow the redirect. If the response is in the error range, the evaluator tool should show a warning that the DMP contains unreachable or erroneous links.

To implement the whitelist for link analysis, one needs to define a list of permitted links for the given field of an maDMP. For example, if an organization allows only a CC-BY license to be assigned to data, then the whitelist must consist of URIs corresponding to different CC-BY licenses. Please note that this is a common requirement to refer to licenses using URIs. For example, the CC-BY license should be linked using <https://creativecommons.org/licenses/by/4.0/>. The tool must map the whitelist to the corresponding field in the maDMP. In the provided example, it would map to `dataset\distribution\license_ref`. The blacklist can be implemented analogously.

The link analysis method described in this section checks whether the links are correct and whether they fulfill specific criteria defined by a funder or an organization. It does not fetch information from the resource the link points to. This is in contrast to methods described in consecutive sections that depend on information retrieved from external resources.

## USING INFORMATION FROM LINKED RESOURCES

Resolving links and fetching additional information from services, registries, and other types of systems to which maDMPs point is crucial in establishing a better context enabling proper assessment of DMPs. Compared to methods in Section 3.2, here we assume that the services in which the linked resources are stored provide custom Application Programming Interfaces (APIs) that we can call to retrieve information on the linked resources. We can distinguish two types of checks:

- **Resource validity:** to check at the specific linked service whether the given resource is valid, e.g. whether a metadata registry has a metadata standard under a specific identifier.
- **Context fetching:** to retrieve additional information on the specific resource that allows the reviewer to better understand the context, e.g. whether the specified metadata standard applies to the domain for which the DMP is created.

We developed a prototype of an evaluator<sup>5</sup> that queries an external resource to get more information on a specific URI used with maDMP. For example, the maDMP specification requires information about the metadata standard that describes the format of the actual metadata of the dataset (`metadata_standard_id`). Instead of creating a new metadata standard, one should link to the existing one from a curated registry. An example of such a registry is FAIRsharing (Sansone et al. 2019). FAIRsharing record for a metadata standard, such as the DataCite Metadata Schema,<sup>6</sup> contains information about license of that standard, its readiness, related taxonomies and domains, and additional metadata.

The evaluator tool takes an URI, for example, `metadata_standard_id`, and checks whether it is registered in `FAIRsharing.org`. If `metadata_standard_id` is verified, it gets a high score due to

---

<sup>5</sup> <https://github.com/vknaisl/madmp-metadata-standard-link-evaluation>.

<sup>6</sup> <https://doi.org/10.25504/FAIRsharing.me4qwe>.

the wide acknowledgment of FAIRsharing. Otherwise, it indicates that this standard is either not suitable/known or that the URI/PID is wrong, and therefore gets penalized with a low score.

In a similar way, a list of suitable/scored URIs or PIDs could be supplied to serve as the evaluation criteria. Furthermore, some of the target metadata about the linked resource could be specified as criteria. For instance, only resources from FAIRsharing that are there marked as ‘ready’ can be used. However, such checks depend on the structure of metadata in the registry and the API provided. Specific constraints can be set by funders or organizations depending on the specific context and requirements.

## USING SCIENTIFIC KNOWLEDGE GRAPHS

Semantic web and specifically Linked Open Data (LOD) (Florian & Martin 2012) allows us to link existing knowledge to help in the interpretation and assessment of DMPs. The semantic web was invented to make machines the primary consumer of information. Similar to Section 3.1, the maDMPs can be serialized using RDF. This in turn enables linking to other concepts from the semantic web and not only makes the information stored in maDMP more precise, but also provides a broader context. For example, for a given dataset, one can fetch additional information on the repository in which it is stored, such as the geographical location or type of the repository. If the repository is located in the United States, then an indicator may be switched reporting a violation of the GDPR regulations for European researchers. The additional information can also reveal that the dataset does not match the domain of the repository, for instance, one should not put biotechnology datasets into a mechanical engineering repository.

One example of an existing Scholarly Knowledge Graph (SKG) (cf. Section 6) is the OpenAIRE Research Graph. We implemented a SPARQL query that fetches additional information on the dataset (see listing 2). The query shows that with a given PID, such as a DOI, we are able to retrieve common metadata as an RDF result set. Many properties are optional in the query. This is necessary since the metadata we wish to retrieve is mostly incomplete. For the example, we see in the first filter statement that we will get metadata like the data source name, subjects, and the title upon execution. However, we will not get any information about their creators or country. Another limitation next to incomplete metadata is the incompleteness of publication records in the OpenAIRE KG. In fact, there is a mismatch between the graphical web browser OpenAIRE Explore and OpenAIRE’s KG. We can find an entry of the publication behind the negative example in our query in OpenAIRE Explore, but not in its KG. However, as this is just a matter of further populating the KG, we see much potential here for future use, especially because RDF is a highly automatable specification and the metadata can therefore be easily linked to maDMPs.

**Listing 2** SPARQL example to query SKG for additional information needed for review of a DMP.

```
SELECT ?pid ?result ?creator ?dateofacceptance ?subject ?project
?resultTitle ?resultType ?country ?datasourceName
WHERE {
  ?result oav:pid | oav:resPersistentID ?pid .
  ?result rdf:type oav:ResultEntity.
  OPTIONAL {?result oav:creator ?creator . }
  OPTIONAL {?result oav:dateofacceptance ?dateofacceptance . }
  OPTIONAL {?result oav:subject ?subject . }
  OPTIONAL {?result oav:outcome ?project . }
  OPTIONAL {?result oav:title ?resultTitle . }
  OPTIONAL {?result oav:resulttype ?resultType . }
  OPTIONAL {?result oav:country ?country . }
  OPTIONAL {?result oav:collectedfrom ?datasourceName . }
  ### Input paramaters: PID/DOI ###
  filter(?pid = "10.5281/zenodo.3974548") # Positive examples
  # filter(?pid = "10.6084/m9.figshare.5718835") # Negative example
}
```

## EXPECTED FAIRNESS ASSESSMENT

Some parts of a data management plan make statements about planned FAIRness which can be then expected from the related results, e.g., published datasets. There are several tools that

support the creation of data management plans including a FAIR metrics evaluation based on users' answers. We can explore this knowledge in the DMP tools and apply it to the maDMP evaluation to assess expected FAIRness and possibly other metrics relevant to a funder.

For example, in the maDMP specification, there is a field indicating the license for a dataset. This field can contain a known license URL, e.g., CC-BY,<sup>7</sup> or any other URL, such as a link to a custom PDF. We can define a list of known URLs that match each of the cases. Then, we can easily evaluate the reusability by confronting answers from a maDMP with such metric-assigned lists based on the knowledge extracted from DMP tools. The individual scoring of answers and their mapping to maDMPs again depends on the specific context and requirements of funders or institutions.

An example of such a tool is the Data Stewardship Wizard (Pergl et al. 2019). Researchers answer questions in a so-called 'smart questionnaire'. FAIR metrics are set for specific answers to indicate whether they are good or bad for a given metric. There is a Common DSW Knowledge Model (capturing the questionnaire structure including metrics configuration) for general data management planning. We can use the FAIR metrics encoded there, and automatically evaluate the FAIRness of certain aspects of maDMP.

With respect to the reusability and license example above, there is the following question about dataset distributions in the model with the answers indicating the *Reusability* metric:

Under what license will the dataset be made available?

- a) They will be freely available for any use (public domain or CC0) **(Reusability = 1)**
- b) They will be freely available with the obligation to quote the source (e.g., CC-BY) **(Reusability = 0.9)**
- c) They will be available under some restrictions **(Reusability = undefined)**

We can explore further questions in the Common DSW Knowledge Model and map them to maDMP fields to implement evaluation metrics. Furthermore, the assessment can be based on knowledge captured in multiple tools, guidelines, best practices, and other resources.

## ACHIEVED FAIRNESS ASSESSMENT

Shortly after the FAIR principles (Wilkinson et al. 2016) were introduced, several tools emerged to address automatic checking of compliance with the principles. Such tools are commonly called FAIR evaluation tools or FAIRness evaluators. Usually a PID of a resource (e.g., DOI, Handle, or URL) is accepted as input, a series of checks for each principle is performed, and results are presented.

An maDMP contains dataset identifiers (*dmp/dataset/dataset\_id/identifier*) that can be used as input for the FAIRness assessment. Thus, for each dataset, an automatic check via API of a FAIRness evaluator can be done. The result can be used for indication (show the score) or alternatively a criterion can be set (e.g., minimal score or required checks to pass).

One of the tools that we can use for the evaluation of FAIRness is the F-UJI automated FAIR data assessment tool (Devaraju & Huber 2022). It is an open-source project and can be easily deployed with a well-documented and easily usable REST API. Then, for each dataset (and its identifier), an HTTP POST request to the */evaluate* endpoint will be made with the identifier in the payload. The response contains detailed information for 16 tests performed, including summarized scoring, which can be presented to the user who requests a DMP evaluation.

The evaluation of FAIRness using F-UJI is prototyped<sup>8</sup> as a simple command-line utility in Python. It takes a DMP in JSON as an input, extracts all dataset identifiers, and tries to evaluate them via request to the configured F-UJI API. The prototype demonstrates the possibility of utilization of external FAIRness evaluators and can be easily further extended, for instance, to have configurable criteria or use detailed results.

Other current and future FAIRness evaluators that provide usable APIs can be integrated and used to score various PIDs from maDMPs. We also investigated FAIRshake (Clarke et al. 2019)

---

<sup>7</sup> <https://creativecommons.org/licenses/by/4.0/>.

<sup>8</sup> <https://github.com/MarekSuchanek/madmp-fairness-evaluation>.



and the FAIR Maturity Evaluation Service (Wilkinson et al. 2019). The first one is relatively complex, but it is also supplied with a Python client library. It allows the user to specify metrics, projects, digital objects, and assessments; a maDMP would be turned into a project with a list of digital objects assessed using specified criteria/metrics. For the second tool, there is no suitable API documentation; however, it is available as an open-source project.<sup>9</sup> It could be deployed with its own maturity indicators specified, forming collections, and then evaluate each PID from a maDMP separately.

## MAPPING OF METHODS TO REVIEWER GOALS

This section reports on the evaluation of existing methods for conducting validation on maDMP against the identified goals of reviewers from Section 2.

Table 1 shows the mapping of goals to methods. If the specific method supports the goal, we indicate it with ‘Y’. Otherwise, the cell is empty.

GOAL/METHOD	RDF-BASED	IDENTIFIER ANALYSIS	LINKED RESOURCES	SKGS	EXPECTED FAIRNESS	ACHIEVED FAIRNESS
<b>G1. Completeness</b>	Y	-	-	-	-	-
<b>G2. Feasibility</b>	-	-	Y	Y	-	-
<b>G3. Quality of actions</b>	-	Y	Y	Y	Y	Y
<b>G4. Non-ambiguity</b>	-	-	-	-	-	-

**Table 1** Mapping of reviewers’ goals to methods for automated assessment of maDMPs.

The results show that only the RDF-based validation helps in assessing the *G1. Completeness* of maDMPs. This is because SHACL constraints are meant to validate contents of RDF documents and, in this case, the maDMP is another type of an RDF document. Goal *G2. Feasibility* is supported by using information from linked resources and from SKGs. This is because these two methods provide means to fetch additional information that renders the information from maDMPs in the wider context: this additional information helps reviewers to assess how realistic and feasible the DMP is. The difference between these two methods is only in the technical aspects of how this context is accessed and represented. Goal *G3. Quality of actions* is supported by all methods except for the RDF-based validation. In this case, methods such as Identifier analysis and Achieved FAIRness check the quality of actions performed, while the three others are more relevant in the planning phase to identify whether planned actions meet community standards or specific requirements. The goal *G4. Non-ambiguity* is not supported by any of the methods. This is because all of the methods discussed in this paper deal uniquely with machine-actionable information; language check of nonstructured text was out of the scope of this publication. Yet, in Section 5 we discuss how this goal can be supported.

## DISCUSSION AND FURTHER RESEARCH CHALLENGES

Although there is currently no known approach for a holistic automated DMP evaluation, we reused existing procedures and tools as integral parts of our methods for the evaluation of maDMPs. In a sense, we extended the potential use cases for these procedures and tools.

In this paper we focused on machine-actionable aspects of maDMPs. A limitation of this method is that a considerable amount of information in maDMPs is still unstructured, due to plain text answers and hence hinders automated assessment and evaluation of maDMPs. Consider the following question from the Science Europe Evaluation Rubric template:<sup>10</sup> “What ethical issues and codes of conduct are there, and how will they be taken into account?”. Answers to this type of questions can vary, e.g in length, number of paragraphs, terms used and sentence structure. This is where Information Retrieval (IR) and Natural Language Processing (NLP) techniques would come in handy. Using IR methods, relevant paragraphs to the questions can be extracted from the DMPs. Hence, methods like extractive question answering (Fajcik et al.

<sup>9</sup> <https://github.com/FAIRMetrics/Metrics/tree/master/MetricsEvaluatorCode/Ruby/fairmetrics>.

<sup>10</sup> Section 4c: [https://www.scienceeurope.org/media/4brkxxe5/se\\_rdm\\_practical\\_guide\\_extended\\_final.pdf](https://www.scienceeurope.org/media/4brkxxe5/se_rdm_practical_guide_extended_final.pdf).

2021) and Named Entity Recognition (NER) can help to extract knowledge, i.e. text segments or concepts, from relevant paragraphs that contain answers to the questions. Thereby, the extracted knowledge can be used to semantically enrich the maDMPs. In our example, codes of conduct and ethical issues could be mapped to entries in large-scale KGs like DBpedia or Wikidata. Part of the assessment, like checking off non-critical ethical issues, could this way be automated and the critical ones could be highlighted for further expert-assessment.

Next to automated evaluation/assessment, another benefit and application scenario for RDF-based maDMPs would be template and requirements refinement. Funding Agencies could use analysis and data mining techniques to aggregate the semantified answers and based on that refine requirements and provide helpful information on how to answer the template questions in a more structured manner, by e.g. providing a set of commonly used licenses based on the previously given answers.

## RELATED WORK

This section presents selected related work that puts our research in context and provides pointers to technologies and methods that we based our research on.

### RDF-BASED CONSTRAINT EVALUATORS

Prior to the development of SHACL, several attempts were made to create a constraint validation mechanism using RDF Graph constraint languages or similar methods. Pellet-ICV (Tao et al. 2010) enables users to work within the Closed World Assumption (CWA) and the Weak Unique Name Assumption (Weak-UNA), allowing application developers to combine open world reasoning and closed world constraint validation in a flexible manner.

Another viable option is SPIN. SPIN<sup>11</sup> is a SPARQL-based rule and constraint language for the Semantic Web, a programming language that combines concepts from object-oriented languages, query languages, and rule-based systems to describe constraints within RDF graph data. RDFUnit (Kontokostas et al. 2014) is a test-driven data debugging framework for linked data. RDFUnit is not a constraint language in the traditional sense, but it can be used to validate and improve the quality of RDF graph data.

Another high-level RDF vocabulary for specifying the shape of RDF resources is Resource Shapes.<sup>12</sup> Resource Shapes are made up of RDF triples that an RDF graph is expected to have and a set of integrity constraints that the RDF graph must meet. Boneva et al. (2017) defined ShEx as a schema formalism for describing the topology of an RDF graph. It defines constraints on the admissible neighborhood for nodes of a given type using regular bag expressions.

In contrast to other approaches, Semantic Web Rule Language (SWRL)<sup>13</sup> was designed to be a rule language rather than a constraint language. In practice, however, many users use SWRL to validate constraints.

### SCIENTIFIC KNOWLEDGE GRAPHS

Persistent identifiers for datasets, funders, grants, metadata, contributors, and also for the DMP itself can point to Scientific Knowledge Graphs (SKG). Generally, a knowledge graph can be defined 'as a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities' (Hogan et al. 2021). In a scientific context, knowledge graphs specify an area of interest that includes the actors, documents, and other research outputs and knowledge (Manghi et al. 2021). An example can be the OpenAIRE Research Graph Data Model, which uses the European OpenAIRE infrastructure, and is a model used to describe scientific objects (Manghi et al. 2019).

---

11 <https://www.w3.org/Submission/spin-overview/>.

12 <https://www.w3.org/Submission/shapes/>.

13 <https://www.w3.org/Submission/SWRL/>.

In the last decade, there has been a shift in a mindset on how to assess scientific output. There are initiatives like COARA,<sup>14</sup> DORA,<sup>15</sup> or Hicks et al.'s (2015) that promote the introduction of new qualitative indicators. The new indicators should serve as a complement to the peer review. Furthermore, the indicators aim to better evaluate the actual content of the scientific output. It should also foster the reduction of the inappropriate usage of the Journal Impact Factor or h-Index for assessing the quality of the research output. The quality indicators should encourage the sharing of data and results, open collaboration or contributions to the research ecosystem, and knowledge generation.

## CONCLUSION

In this paper, we discussed methods for automating DMP assessment. To do so, we identified the different goals that reviewers of DMPs may have. These include DMP completeness, feasibility, quality of actions, and non-ambiguity. We also analyzed scenarios in which methods can be used for automated assessment of DMPs. These go beyond the review of the DMP by a funder, and include settings in which the assessment is used as a continuous feedback to the person creating and improving a DMP. We identified methods and provided examples and implementations of them to demonstrate how specific reviewer tasks can be automated. The methods depend on the machine-actionability of DMPs and include semantic web technologies, integration with registries and knowledge graphs, and FAIRness evaluation tools. We mapped the methods to reviewers' goals to identify which goals can be supported using specific methods. The methods investigated in this paper can be applied to different phases of a DMP lifecycle: at the early stages, when DMP is more aspirational, or at the later stages, when DMP mostly describes actions that were already performed.

The proposed methods constitute a toolbox that can be used to build specific tools for automated maDMP assessment. Specific tools must take into account the exact constraints in which the tools will be used, for instance, reflecting specific funder requirements, or institutional policies or legal constraints.

As a follow-up to our work, we plan to map these methods to popular funder templates to identify the level and kind of automation possible. In the long term, we plan to use Information Retrieval and Natural Language Processing (NLP) techniques to create further methods that better address the non-machine-actionable parts of maDMPs.

## FUNDING INFORMATION


The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Program.


## COMPETING INTERESTS

T.M. is a co-editor of the special issue in the CODATA Data Science Journal on "Data Management Planning across Disciplines and Infrastructures". Otherwise the authors have no competing interests to declare.

## AUTHOR AFFILIATIONS

**Tomasz Miksa**  [orcid.org/0000-0002-4929-7875](https://orcid.org/0000-0002-4929-7875)  
SBA Research, AT; TU Wien, AT

**Marek Suchánek**  [orcid.org/0000-0001-7525-9218](https://orcid.org/0000-0001-7525-9218)  
Czech Technical University in Prague, CZ

**Jan Slifka**  [orcid.org/0000-0002-4941-0575](https://orcid.org/0000-0002-4941-0575)  
Czech Technical University in Prague, CZ

---

<sup>14</sup> Coalition for Advancing Research Assessment <https://coara.eu/>.

<sup>15</sup> The Declaration on Research Assessment <https://sfedora.org/>.

Vojtech Knaisl  [orcid.org/0000-0003-0103-8468](https://orcid.org/0000-0003-0103-8468)

Czech Technical University in Prague, CZ

Fajar J. Ekaputra  [orcid.org/0000-0003-4569-2496](https://orcid.org/0000-0003-4569-2496)

TU Wien, AT; Wirtschaftsuniversität Wien, AT

Filip Kovacevic  [orcid.org/0000-0002-2854-0434](https://orcid.org/0000-0002-2854-0434)

TU Wien, AT

Annisa Maulida Ningtyas  [orcid.org/0000-0002-5041-0230](https://orcid.org/0000-0002-5041-0230)

TU Wien, AT

Alaa El-Ebshihy  [orcid.org/0000-0001-6644-2360](https://orcid.org/0000-0001-6644-2360)

TU Wien, AT

Robert Pergl  [orcid.org/0000-0003-2980-4400](https://orcid.org/0000-0003-2980-4400)

Czech Technical University in Prague, CZ

Miksa et al.

*Data Science Journal*

DOI: 10.5334/dsj-2023-

028

12

## REFERENCES

- Boneva, I, Labra Gayo, JE, Prud'hommeaux, EG. 2017. Semantics and Validation of Shapes Schemas for RDF. In: d'Amato, C., et al. (eds.), *The Semantic Web – ISWC 2017. ISWC 2017. Lecture Notes in Computer Science*, 10587. Cham: Springer. DOI: [https://doi.org/10.1007/978-3-319-68288-4\\_7](https://doi.org/10.1007/978-3-319-68288-4_7)
- Cardoso, J, Castro, LJ, Ekaputra, FJ, Jacquemot, MC, Suchanek, M, Miksa, T and Borbinha, J. 2022. DCSO: Towards an ontology for machine-actionable data management plans. *Journal of Biomedical Semantics*, 13(21). DOI: <https://doi.org/10.1186/s13326-022-00274-4>
- Clarke, D, Wang, L, Jones, A, Wojciechowicz, M, Torre, D, Jagodnik, K, Jenkins, S, McQuilton, P, Flamholz, Z, Silverstein, M, Schilder, B, Robasky, K, Castillo, C, Idaszak, R, Ahalt, S, Williams, J, Schurer, S, Cooper, D, de Miranda-Azevedo, R, Klenk, J, Haendel, M, Nedzel, J, Avillach, P, Shimoyama, M, Harris, R, Gamble, M, Poten, R, Charbonneau, A, Larkin, J, Brown, C, Bonazzi, V, Dumontier, M, Sansone, SA and Ma'ayan, A. 2019. Fairshake: Toolkit to evaluate the fairness of research digital resources. *Cell Systems*, 9(5): 417–421. DOI: <https://doi.org/10.1016/j.cels.2019.09.011>
- Devaraju, A and Huber, R. 2022. *F-ujj – An automated fair data assessment tool (v1.0.0)*. Available at <https://zenodo.org/record/4063720> [Last accessed 03 August 2023].
- Science Europe. 2021. *Practical Guide to the International Alignment of Research Data Management – Extended Edition*. Available at <https://zenodo.org/record/4915862> [Last accessed 03 August 2023].
- Fajcik, M, Jon, J and Smrz, P. 2021. Rethinking the objectives of extractive question answering. In: *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*. Punta Cana, Dominican Republic, Association for Computational Linguistics, pp. 14–27. November: <https://aclanthology.org/2021.mrq-a-1.2>.
- Florian, B and Martin, K. 2012. *Linked Open Data: The Essentials – A Quick Start Guide for Decision Makers*. Vienna: Ed. mono/monochrom.
- Foidl, R, Brugger, LS and Miksa, T. 2021. Automating Evaluation of Machine-Actionable Data Management Plans with Semantic Web Technologies. In: *DaMaLOS – 2nd Workshop on Data and Research Objects Management for Linked Open Science, Co-located at the International Semantic Web Conference ISWC 2021*. PUBLISSO, 24 October 2021, pp. 1–13. DOI: <https://doi.org/10.4126/FRL01-006429413>
- Hicks, D, Wouters, P, Waltman, L, de Rijcke, S and Rafols, I. 2015. Bibliometrics: The leiden manifesto for research metrics. *Nature News*, 520(7548): 429. DOI: <https://doi.org/10.1038/520429a>
- Hogan, A, Blomqvist, E, Cochez, M, D'Amato, C, Melo, GD, Gutierrez, C, Kirrane, S, Gayo, JEL, Navigli, R, Neumaier, S, Ngomo, ACN, Polleres, A, Rashid, SM, Rula, A, Schmelzeisen, L, Sequeda, J, Staab, S and Zimmermann, A. 2021. Knowledge graphs. *ACM Comput. Surv*, 54(4): 1–37. DOI: <https://doi.org/10.1145/3447772>
- Kontokostas, D, Westphal, P, Auer, S, Hellmann, S, Lehmann, J, Cornelissen, R and Zaveri, A. 2014. Test-driven evaluation of linked data quality. In: *Proceedings of the 23rd International Conference on World Wide Web. WWW'14*, Association for Computing Machinery, New York, NY, pp. 747–758. DOI: <https://doi.org/10.1145/2566486.2568002>
- Manghi, P, Bardi, A, Atzori, C, Baglioni, M, Manola, N, Schirrwagen, J and Principe, P. 2019. *The openaire research graph data model (1.3)*. Available at: <https://zenodo.org/record/2643199> [Last accessed 03 August 2023].
- Manghi, P, Mannocci, A, Osborne, F, Sacharidis, D, Salatino, A and Vergoulis, T. 2021. New trends in scientific knowledge graphs and research impact assessment. *Quantitative Science Studies*, 2(4): 1296–1300. DOI: [https://doi.org/10.1162/qss\\_e\\_00160](https://doi.org/10.1162/qss_e_00160)
- Miksa, T, Oblasser, S and Rauber, A. 2021. Automating research data management using machine-actionable data management plans. *ACM Trans Manage Inf Syst*, 13(2): 1–22. DOI: <https://doi.org/10.1145/3490396>

- Miksa, T, Simms, S, Mietchen, D and Jones, S.** 2019. Ten principles for machine-actionable data management plans. *PLoS computational biology*, 15(3): e1006750. DOI: <https://doi.org/10.1371/journal.pcbi.1006750>
- Miksa, T, Walk, P and Neish, P.** 2020. *RDA DMP common standard for machine-actionable data management plans*. Available at: <https://doi.org/10.15497/rda00039> [Last accessed 03 August 2023].
- Miksa, T, Walk, P, Neish, P, Oblasser, S, Murray, H, Renner, T, Jacquemot-Perbal, MC, Cardoso, J, Kvamme, T, Praetzelis, M, Suchánek, M, Hooft, R, Faure, B, Moa, H, Hasan, A and Jones, S.** 2021. Application profile for machine-actionable data management plans. *CODATA Data Science Journal*, 20(1): 32. DOI: <https://doi.org/10.5334/dsj-2021-032>
- Pergl, R, Hooft, R, Suchánek, M, Knaisl, V and Slifka, J.** 2019. 'Data stewardship wizard': A tool bringing together researchers, data stewards, and data experts around data management planning. *Data Science Journal*, 18(1): 59. DOI: <https://doi.org/10.5334/dsj-2019-059>
- Sansone, S, McQuilton, P, Rocca-Serra, P, Gonzalez-Beltran, A, Izzo, M, Lister, A, Thurston, M and Community, F.** 2019. Fairsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, 37(4): 358–367. DOI: <https://doi.org/10.1038/s41587-019-0080-8>
- Tao, J, Sirin, E, Bao, J and McGuinness, DL.** 2010. Integrity constraints in owl. In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI'10, AAAI Press, Atlanta Georgia, July 11–15, 2010, pp. 1443–1448. DOI: <https://doi.org/10.1609/aaai.v24i1.7525>
- Wilkinson, MD., Dumontier, M, Aalbersberg, IJ., Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, JW, da Silva Santos, LB, Bourne, PE, et al.** 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 15(3): 160018. DOI: <https://doi.org/10.1038/sdata.2016.18>
- Wilkinson, MD, Dumontier, M, Sansone, SA, Olavo, L, Prieto, M, Batista, D, McQuilton, P, Kuhn, T, Rocca-Serra, P, Crosas, M and Shultes, E.** 2019. Evaluating fair maturity through a scalable, automated, community-governed framework. *Nature-Springer Scientific Data*, 6(174). DOI: <https://doi.org/10.1038/s41597-019-0184-5>

**TO CITE THIS ARTICLE:**

Miksa, T, Suchánek, M, Slifka J, Knaisl V, Ekaputra FJ, Kovacevic F, Ningtyas AM, El-Ebshihy, A and Pergl, R. 2023. Towards a Toolbox for Automated Assessment of Machine-Actionable Data Management Plans. *Data Science Journal*, 22: 28, pp. 1–13. DOI: <https://doi.org/10.5334/dsj-2023-028>

**Submitted:** 15 December 2022

**Accepted:** 25 May 2023

**Published:** 28 August 2023

**COPYRIGHT:**

© 2023 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.