Towards Semantic Data Management Plans for Efficient Review Processing and Automation

Jana Martínková[a] and Marek Suchánek[b]
Faculty of Information Technology, CTU in Prague, Prague, Czech Republic

Keywords: Data Management Plan, Semantic Annotation, Ontology, Machine-Actionable, Human-Readable, DMP Template.

Abstract: In recent times, Data Management Planning has become increasingly crucial. Effective practices in data management ensure more precise data collection, secure storage, proper handling, and utilization beyond the primary project. However, existing DMPs often suffer from complex structures that impede accessibility for humans and machines. This project aims to address these challenges by converting DMPs into formats that are both machine-actionable and human-readable. Leveraging established DMP templates and relevant ontologies, our methodology involves analyzing diverse approaches to achieve this dual functionality. We assess machine-actionability through comparative evaluations using AI and NLP tools. Furthermore, we identify gaps in ontologies, laying the groundwork for future enhancements in this critical area of research.

1 INTRODUCTION

The primary objective of this project is to propose a way of capturing Data management plans (DMPs) that are both machine-actionable and human-readable. This goal aspires to bridge the gap between conventional review procedures and the potential for automation. Unfortunately, current DMPs often suffer from convoluted composition, which impedes their accessibility to human readers, and their structure fails to align with machine-friendly processing. In recognition of this challenge, our proposal focuses on the development of a DMP template that harmoniously combines a human-readability and machine-actionability while considering the existing work that has been done in terms of funder templates as well as machine-actionable DMPs specification. To accomplish this, we set the following partial steps:

- **G1.** find and review 3 suitable DMP templates
- **G2.** manually annotate the parts/questions from each selected DMP templates using ontologies and vocabularies related to the DMP or defined suitable terms
- **G3.** use formats combining the machine-actionability and human-readability to capture existing DMP enriched with semantic annotations
- **G4.** evaluate the usability and correctness of the proposed solutions

2 MACHINE-ACTIONABLE VS. HUMAN-READABLE DOCUMENTS

Machine-readable and human-readable documents each serve distinct purposes, necessitating unique characteristics, structures, and potential technologies for their creation and interpretation. Machine-readable documents, designed primarily for automated processing by computers, prioritize structured data, often employing formats like XML, JSON, or CSV, enabling seamless data extraction and analysis (Open Knowledge, 2015).

In contrast, human-readable documents cater to human comprehension and are typically presented in a visually appealing format with rich content, including text, images, and multimedia elements.

On the other hand, machine-actionable data, as described in (Data Documentation Initiative, 2023), refers to structured data that machines or computers can be programmed against its structure. Moreover by (ELIXIR, Research Data Management Kit, 2021),

[a] https://orcid.org/0000-0001-8575-6533
[b] https://orcid.org/0000-0001-7525-9218

Martínková, J. and Suchánek, M.
Towards Semantic Data Management Plans for Efficient Review Processing and Automation.
DOI: 10.5220/0012837900003756

In *Proceedings of the 13th International Conference on Data Science, Technology and Applications (DATA 2024)*, pages 543-550
ISBN: 978-989-758-707-8; ISSN: 2184-285X

machine-actionable data fosters semantic and syntactic data integration, particularly among datasets sharing similar experimental conditions or variables.

There are formats that try to combine the capabilities of machine processing without losing human capabilities. These include, for example, RDFa or so-called microformats.

2.1 RDFa

Resource Description Framework in Attributes (RDFa) (RDFa Working Group, 2013) is a W3C recommendation that enables the representation of structured data by utilizing attributes in Hyper Text Markup Language (HTML) elements. By incorporating RDFa into web pages, it becomes possible to embed semantic annotations while maintaining the human-readable content of the web page.

2.2 Microformats

Microformats (Khare, 2006) are a collection of predefined HTML classes that empower data formats to be enriched with semantic meaning. These HTML-classes allow for the inclusion of machine-actionable data directly in web page content.

2.3 Microdata

Microdata (Web Hypertext Application Technology Working Group, 2023) share similarities with Microformats, but a key distinction lies in their generality. Unlike Microformats, Microdata is not predefined, providing greater flexibility by allowing the use of various ontologies and vocabularies according to the author's preferences. This makes Microdata well-suited for scenarios requiring the annotation of diverse data within a single document.

2.4 XSL Transformation

XML (eXtensible Markup Language) (Bray et al., 2008) plays a pivotal role as a versatile format for both machine-readable and human-readable documents. While primarily designed for machine readability, XML's intuitive markup is also accessible to humans, especially when properly formatted.

XSL (eXtensible Stylesheet Language) (W3C, 2017) complements XML by enabling the transformation of machine-readable XML data into more human-readable formats. XSLT (XSL Transformations) (Kaz, 2017) is a key component of XSL, providing a powerful mechanism for converting XML documents into different output formats, such as HTML, PDF, or plain text.

XSD (XML Schema Definition) (W3C, 2012) plays a crucial role in ensuring the integrity and validity of XML documents. XSD provides a set of rules and constraints that define the structure and data types within an XML document. This schema validation maintains data consistency and reliability, indirectly enhancing human readability.

3 DATA MANAGEMENT PLANS

DMP is a document that facilitates efficient data management throughout a project. It outlines the lifecycle of the data created or collected during the project, detailing how the data will be handled and providing information about their future usability and availability. Effective data management practices lead to more accurate data collection, secure storage, and proper handling, elevating their potential value and relevance in diverse research domains. (Smale et al., 2018)

A DMP is commonly structured using a standardized template to ensure all essential components are covered, although certain sections may be adapted based on the project, funding source, or organization. In this work, the Horizon Europe, Science Europe, and National Institutes of Health (NIH) DMP templates are described in more detail to fulfill the G1 goal. These templates were selected due to their extensive adoption on a global scale.

Science Europe Template (Science Europe, 2021) includes essential details and a table that links different DMP sections with individual Findable, Accessible, Interoperable, Reusable (FAIR) principles.

Horizon Europe Template (European Commission, 2020) covers all the necessary parts of the data knowledge and includes questions explicitly aligned with the FAIR principles, organized according to the structure of those principles.

National Institutes of Health Data Management and Sharing Plan Template (National Institutes of Health, 2023) is very simple, which can lead to insufficient information being filled in. It completely omits the connection with the FAIR principles and furthermore does not cover the areas of legal requirements, data storage security and allocated resources for data management in the project. On the other hand, its brevity should not discourage the data steward from completing this plan at the beginning.

In order to obtain all the necessary information, the *National Institutes of Health Data Management and Sharing Plan Template* (National Institutes of Health, 2023) is insufficient. On contrast, the *Sci-*

ence Europe Template (Science Europe, 2021) and the *Horizon Europe Template* (European Commission, 2020) cover all the crucial details required by the DMP. From the user's perspective, the *Science Europe Template* (Science Europe, 2021) provides a more pleasant experience, as its questions do not place strong emphasis on the FAIR principles. Instead, they prompt the data steward to consider how they approach various issues within the project, rather than immediately focusing on applying the specific FAIR principle in question.

4 RELATED WORK

According to (DataCite, 2021) Machine-actionable data management plans (maDMPs) play a pivotal role in fostering the exchange of information by establishing connections between metadata and various sources, including repositories and institutions.

The primary objective of the DMP Common Standard (DCS) Working Group, under the purview of Research Data Alliance (RDA), centres on the establishment of well-defined processes for research data management, a robust data management infrastructure, and, most importantly, a universally accepted standard (Miksa et al., 2021) in the form of a data model to represent DMP information. Its implementation ensures seamless interoperability between systems engaged in producing or consuming maDMP, while concurrently permitting the assimilation of additional information from diverse systems, such as organizational or repository-related data. Within the framework of the DCS Working Group's endeavours, a JSON serialization of this application profile has been generated, offering practical utility.

However, as noted by (Cardoso et al., 2022), several aspects of this profile present challenges. Foremost among them is the absence of direct, explicit linkages to existing ontologies or vocabularies. Additionally, DCS covers only essential parts of the DMPs, omitting crucial elements such as the provenance of reused or generated data, project objectives, data access embargo, or the access protocol.

While the profile is designed to be extensible, a discernible mechanism for segregating the foundational specification from its extensions is not yet defined. Pertinently, while the profile aims for machine-actionability, limitations arise due to certain elements within the structure accommodating text fields.

The DMP Common Standard Ontology (DCSO) (Cardoso et al., 2022), an ontology grounded in the RDA's DCS standard specification, addresses these concerns.By consolidating terms into a comprehensive ontology intertwined with several pre-existing ontological constructs, the DCSO significantly enhances interoperability within the realm of the RDA DCS working group standard.

One of the assessments of the DCSO is conducted through the maDMP Evaluation (Foidl and Burgger, 2021). This investigation involves the transformation of openly available maDMP instances into DCSO instances using the *dcsojon* tool, generating JavaScript Object Notation for Linked Data (JSON-LD) representations. By applying predefined SPARQL Protocol and RDF Query Language (SPARQL) (Harris and Seaborne, 2013) queries based on the evaluation methodology outlined in *the International Alignment of Research Data Management* (Science Europe, 2021) DMP template, the study assesses the expressiveness of the SPARQL queries against the evaluation rubric criteria.

The challenges in expressing certain criteria via SPARQL queries often stem from the fact that the concept is either not covered or only partially covered by the fundamental DCSO, thus remaining absent in the transformed DMP DCSO.

5 OUR APPROACH

To accomplish the defined objectives of this study, we initiated the annotation of existing DMP instances with terms extracted from diverse ontologies and dictionaries, as elaborated in Section 5.1, aligning with the fulfilment of G2. After these annotations, we proceeded to implement them in various formats, thoroughly evaluating their advantages and drawbacks, as detailed in Section 5.2, thereby achieving G3. Finally, the manually implemented maDMP underwent comprehensive assessment through multiple approaches, as outlined in Section 6, to meet the objectives of G4.

5.1 Annotations

During the development of a machine-actionable and human-readable DMP, terms from various ontologies and dictionaries were gradually grouped to provide a sufficient semantic description of the information contained in the DMP. Throughout the annotation process, approximately nine common ontologies and vocabularies were employed. Even in this initial phase, it became apparent that not all known existing ontologies and dictionaries were sufficient. For this reason, problematic areas are described below, along with proposals on how to address them. In connection with term proposals for annotations, a fictional non-existent ontology with the prefix "dmp:" was created.

Terms in the DMP with this prefix represent proposed terms that could be used for annotation. However, they are not properly defined.

Furthermore, existing DMP were manually annotated. It was found that although the wording of questions in DMP templates may be unambiguous, the answers do not always contain or cannot contain the desired information. As a result, while the annotations of individual existing DMPs are based on grouped terms from the previous step, in some cases, they had to be adjusted or supplemented to correspond to the information in the existing DMPs, thus providing a true semantic description.

5.1.1 Core Elements of a DMP

The core constituents of the DMP encompass the DMP document itself, the relevant project, the ensuing data, along with its metadata, and the designated repository for data storage. Given that a significant portion of the information within the DMP is interlinked mainly with these components, it is crucial to accurately define and delineate these elements.

Figure published online[1] (Martínková and Suchánek, 2024) illustrates the annotations of the fundamental components of the DMP together with their types, indicating the class to which each individual or node belongs. The *typeof* relationship is represented by a dashed arrow, the object property relationship is depicted by a solid arrow. Some core elements are *typeof* more than one class to meet the requirements of object or data properties and inheritance specifications in the term definitions.

The following section addresses the primary challenges and intriguing aspects associated with annotating the DMP.

5.1.2 Information About Reusing Datasets

Within the context of the DMP, it is vital to ascertain whether any datasets are reused throughout the research process. While the DCSO does not provide an explicit solution, the *dcso:Dataset* can potentially have a defined creation date before the project's inception, which can indicate that the dataset is reused in the current project. This solution is not very intuitive, nevertheless in the commonly used ontologies there is no suitable solution.

Several solutions were considered for this issue. The initial approach involved employing the object property *prov:wasDerivedFrom* to establish a link between the resulting dataset and the reused dataset. However, it is often necessary to explain the rationale for reuse and specify if certain datasets are ultimately unused. This information cannot be captured solely by the *prov:wasDerivedFrom* property.

Our approach uses the *dcat:qualifiedRelation* property to link a dataset to a *dcat:Relationship* instance, which connects to the reused dataset via *dcat:relation*. The *dmp:reason* property provides the justification for reuse, and an element can be added for discarded reuse cases.

Furthermore, the data property *dmp:reusingData* has been introduced. Although this may seem like duplication of information capture, it serves for instances where there are no reused datasets or the information is unknown. As a result, it captures only the values *Yes*, *No*, or *Unknown*.

5.1.3 Information About the Purpose of the Data and Its Relation to Objectives of the Project

In the DMP is usually captured the purpose of the resulting data together with its relation to the objectives of the project. However the DCSO doesn't cover this aspect and there are no terms for capturing objectives and their relations in common ontologies.

Hence, additional terms were introduced into the hypothetical *dmp* ontology. The object property *dmp:hasObjective* with a range of *rdfs:Resource* was established to express a project's objective. To link a dataset with the project objective, the object property *dmp:fulfillsObjective* was also added.

5.1.4 Metadata Elements

When detailing the metadata associated with datasets within the DMP, it is crucial to explicitly specify the used metadata schema or the individual metadata elements employed. In the case of the former, well-known ontologies adequately address this requirement but expressing individual metadata elements becomes challenging. To overcome this limitation, the data property *dmp:containsMetadataElement* was introduced to express the individual metadata elements.

5.1.5 Information About the Trustworthiness of the Data Repository

In the DMP the trustworthiness of the used data repository(ies) are usually captured. Unfortunately, prevalent ontologies lack suitable solutions to encapsulate the trustworthiness. In practice, repositories establish trustworthiness by adhering to the TRUST principles (Lin et al., 2020) or obtaining dedicated certificates. Unfortunately, the DMP do not usually specifically ask for the reason for trustworthiness. In such

[1] https://doi.org/10.5281/zenodo.10893770

cases, information could be easily annotated using an object property linking the repository to the trust certificate or the evaluation of TRUST principles. However, responses in the DMP commonly only state "Yes, the repository we use is trusted".

To address this, the data property *dmp:isTrusted* was introduced to describe the repository with values restricted to "Yes," "No," or "Unknown."

5.1.6 Availability and Accessibility of Data and Metadata

Within the DMP, multiple inquiries focus on the availability and accessibility of both data and associated metadata. Upon examining various DMPs, it becomes apparent that researchers often provide similar answers or similar key ideas to these questions. Some questions cover multiple aspects, making it challenging to comprehensively address each point and resulting in insufficient DMP outcomes.

In this critical aspect of DMPs, it would be beneficial to explore a more structured approach to obtain this valuable information, perhaps even incorporating semi-controlled vocabulary options in certain sections. To annotate this area a whole new approach was designed as shown in the figure published online[2] (Martínková and Suchánek, 2024). It's important to bear in mind that this is just an initial proposal. The primary goal of this work was not to create a new ontology for DMP, but during this work, several deficiencies in the existing options were identified.

To facilitate annotation, we established classes and properties within a hypothetical DMP ontology. The design of this structure primarily revolves around defining two key concepts: availability and accessibility, in alignment with the *Common DSW Knowledge Model* (DSW Team, 2018) used in the Data Stewardship Wizard (DSW) (Pergl et al., 2019).

5.1.7 Common Ontologies and Vocabularies

In comprehensive DMP templates, the question arose about using ontologies and vocabularies in the data context. Since a specific term for ontology or vocabulary wasn't found, a concept was created in the hypothetical *dmp:* ontology, as can be seen in the figure published online[2] (Martínková and Suchánek, 2024).

5.1.8 Cost and Its Funding

The DMP typically addresses the resources necessary to meet FAIR principles. For quantifying the resources allocated to make data FAIR, we employed the object property *schema:estimatedCost* with a range of *schema:MonetaryAmount*, specifying value and currency. To signify that this amount pertains to enhancing data FAIRness, we utilized the *sioc:Topic*.

To annotate how these expenditures will be funded, the object properties *schema:funder* and *foaf:fundedBy* were employed.

[2]https://doi.org/10.5281/zenodo.10893770

5.2 Different Formats

In this study, three potential formats, as described in Section 2, were considered to capture annotations within a human-readable environment DMPs. This section details their capabilities and assesses their suitability for the intended purpose.

5.2.1 XSL Transformations

Utilizing Extensible Markup Language (XML) for annotation and subsequent transformation to human-readable text using Extensible Stylesheet Language Transformations (XSLT) is not entirely suitable for our requirements. The main problem is with the organization of DMPs. Questions are often grouped by topic rather than by core elements, resulting in scattered references to datasets throughout the document.

Code Example 1: Example of XML annotations.

```
<dataset>
  This data set has following distributions:
  <distributions>
    <distribution>
      <title>Distribution A</title>
      has size
      <bytesize>10 MB</bytesize>
...
    <availability> Yes, all data will be
    ↪  made openly available.
```

However, XML is structured and once an element is in the document, it cannot be repeated. In example 1, there are two *dataset* elements, even though, in the context of the DMP, we are referring to only one dataset. This limitation led us to explore alternative approaches rather than continuing with this method.

5.2.2 Microformats and Microdata

Microformats prove to be unsuitable for our work due to the diverse nature of information within the DMP. On the contrary, Microdata is highly suitable, offering the flexibility to use any ontology or vocabulary for semantic annotations. However, Microdata cannot create Resource Description Framework (RDF) blank nodes. Since various elements in the DMP lack properly defined identifiers but are referenced in multiple sections, blank nodes become valuable. They enable

the aggregation of information about, for example, a reused data set mentioned in different parts of the entire DMP. Even without explicit identifiers, the use of blank nodes allows the connection of information related to the same dataset.

5.2.3 RDFa

RDFa emerges as the optimal solution for our work, offering the flexibility to utilize any ontology or vocabulary for text annotations. It also facilitates the creation of blank nodes, allowing the connection of information to these nodes via object or data properties throughout the document by assigning them a node identifier within the local resource.

Code Example 2: Example of RDFa annotations.

```
<b>Will you re-use any existing data?</b>
<span resource="#dataset"
↪  typeof="dcat:Dataset schema:Thing">
<span property="dmp:reusingData">No </span>
↪  data will be re-used.
```

In the code example 2, the previously defined *dataset* as a named blank node allows for contextual connections, linking all information related to the *dataset* irrespective of its position within the document. This leads to the choice of RDFa as the preferred format for annotations in our work.

6 EVALUATION

The assessment in this study involved comparing semantically annotated DMPs against Natural Language Processing (NLP) and text-mining methods, including ChatGPT, applied to human-readable (non-annotated) DMPs. A foundational list of approximately 20 questions, covering all aspects of the DMP, served as the benchmark for testing these approaches on both annotated and non-annotated DMPs. The annotated DMPs and the list of evaluation questions are published online[3] (Martínková and Suchánek, 2024).

The evaluation involved two distinct methodologies applied to non-annotated DMPs. Initially, the open-source NLP tool, the Hugging Face Transformer (Wolf et al., 2020), and ChatGPT (OpenAI, 2022) were utilized. Each tool was tasked with querying the DMPs using questions in natural human language, and the provided responses were manually evaluated comparing the information contained in the DMP. For semanticaly annotated DMPs, SPARQL queries aligned with the set of questions were utilized.

A comprehensive evaluation was conducted on a total of 5 DMPs: 2 following the *Horizon Europe* template (European Commission, 2020), 2 adhering to the *NIH Data Management and Sharing Plan* (National Institutes of Health, 2023), and just one DMP aligning with the *International Alignment of Research Data Management* (Science Europe, 2021). The selection of only one DMP in the latter case stems from the observed discrepancy between the DMPs available online and their original templates, rendering them unsuitable for the evaluation process. The small number of DMPs is due to the fact that performing detailed annotation manually is a lengthy process; nevertheless, we still managed to achieve results.

The table 1 displays the percentage of correctly answered questions for each method and each individual DMP. Initially, the proportion of 20 evaluation questions that could be answered for each DMP, indicating the presence of this specific information in the DMP, was determined. All results from the evaluation methods are calculated based on this percentage, not the total of 20 questions.

6.1 Evaluation Questions

A set of approximately 20 questions was devised to encompass aspects related to the reuse, resulting data, metadata, their availability and accessibility, resources for ensuring FAIRness, and legal and ethical considerations. To assess annotated DMPs, corresponding SPARQL queries were formulated. The full list of evaluation question is published online (Martínková and Suchánek, 2024).

6.2 Hugging Face Transformer

The Hugging Face Transformer (Wolf et al., 2020) provides a framework and pre-trained models, simplifying the performance of the NLP and especially for our work the method of Question answering.

We used the Disilbert model (Sanh et al., 2019) to analyze unannotated DMPs by posing questions from the list in natural human language. However, the tool's answers were highly inadequate. The tool often generated entirely unreasonable responses. The table 1 in the column labeled "The Hugging Face Transformer" shows the percentages of cases where the answer closely aligned with the queried topic. One question that consistently received accurate responses pertained to the volume of resulting data.

[3]https://doi.org/10.5281/zenodo.10893770

Table 1: Percentage of accurately answered questions for each method and each individual DMP.

	Template	Answerable	The Hugging Face Transformer	ChatGPT	SPARQL
DMP1	Horizon Europe	71.43%	33.33%	86.67%	100.00%
DMP2	Horizon Europe	100.00%	9.52%	90.48%	100.00%
DMP3	Science Europe	61.90%	15.38%	76.92%	100.00%
DMP4	NIH	47.62%	20.00%	70.00%	100.00%
DMP5	NIH	33.33%	42.86%	71.43%	100.00%

6.3 ChatGPT

The ChatGPT is a language (OpenAI, 2022) model, designated for natural language understanding an generation. This model was utilized for the analysis of unannotated DMPs by formulating questions from the list in natural human language. The tool yields highly accurate and detailed responses, achieving very high correctness, and the answers can be serialized in various formats upon request. The precision of the responses to the evaluation questions is shown in the table 1 under the column labeled "ChatGPT."

However, there are drawbacks to this solution. First, it tends to provide lengthy responses, including surrounding context. While this can be seen as an advantage depending on specific requirements, for machine-actionable purposes, simplicity with a comprehensive description is preferable.

The second limitation is that the tool lacks the capability to count automatically. If the DMP includes a list of reused datasets, the tool cannot provide the exact number of reused datasets. It's important to note that while ChatGPT occasionally struggles with poorly described and ambiguous text, this issue is typically less prominent in the context of the DMP.

6.4 SPARQL

SPARQL (Harris and Seaborne, 2013) is a standardized query language for retrieving and manipulating data in RDF structure. SPARQL queries were used for the analysis of annotated DMPs by formulating questions as SPARQL queries. Not surprisingly, this approach was the most sufficient and provided the exact answers (if they were part of DMP) as can be seen in the table 1 under the column "SPARQL."

The incorporation of pre-existing annotations helps prevent misunderstandings in complex, semistructured, or intuitive texts where ChatGPT might otherwise encounter challenges. Contrary to ChatGPT, the utilization of SPARQL allows counting operations. Therefore, assessing the number of reused datasets, the total volume of datasets in various units or the sum of required resources can be conveniently obtained.

7 CONCLUSION

Three DMP templates were chosen and annotated based on the semantic meaning of their parts and questions, aligning with the objectives **G1** and **G2**.Nine well-known ontologies, including DCSO were utilized alongside the introduction of additional terms when needed. Over 20 terms were specifically defined.

The identified gaps among known ontologies within the DMP domain present future opportunities to enhance the proposed solution and provide more comprehensive coverage. The analysis involved capturing these annotations in different formats, with RDFa chosen to represent all five annotated DMP examples, fulfilling objective **G3**.

To fulfill objective **G4**, a manual evaluation of usability and correctness took place, involving a total of 20 questions. The usability and correctness of the proposed solutions were assessed by testing unannotated DMP instances without their semantic meaning using two approaches: The Hugging Face Transformer (Wolf et al., 2020) and ChatGPT (OpenAI, 2022). However, both approaches did not achieve as high percentage of correctly answered questions as the approach using equivalent SPARQL queries on semantically annotated DMPs.

It is important to note that semantically annotated DMPs contain ontological terms matching the SPARQL version of the evaluation questions. Because the infromation caputerd within the DMPs and evaluation question pertains to the same domain. These results demonstrate the usability of this approach, strongly suggesting that combining manual annotation with NLP or Artificial intelligece (AI) methods could streamline the process, making it intriguing avenue for future exploration.

ACKNOWLEDGEMENTS

This work was supported by the Student Summer Research Program 2023 of the FIT CTU in Prague and by the Czech Technical University in Prague grant: Advance Research In Software Engineering, No. SGS23/206/OHK3/3T/18.

REFERENCES

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F. (2008). Extensible markup language (xml) 1.0 (fifth edition). [online]. [Accessed 2023-08-13].

Cardoso, J., Castro, L. J., Ekaputra, F. J., Jacquemot, M. C., Suchánek, M., Miksa, T., and Borbinha, J. (2022). DCSO: towards an ontology for machine-actionable data management plans. *Journal of Biomedical Semantics*, 13(1):21.

Data Documentation Initiative (2023). Machine-actionable. [Accessed 2023-07-17].

DataCite (2021). Introduction to machine actionable dmps (madmps). [online]. [Accessed 2023-08-13].

DSW Team (2018). Common DSW Knowledge Model. [online]. [Accessed 2023-03-19].

ELIXIR, Research Data Management Kit (2021). Machine-actionability. [Accessed 2023-07-17].

European Commission (2020). Horizon 2020 dmp. [online]. [Accessed 2023-12-15].

Foidl, R. and Burgger, L. S. (2021). Evaluation of maDMPs using SPARQL.

Harris, S. and Seaborne, A. (2013). Sparql 1.1 query language. [online]. [Accessed 2023-08-13].

Kaz, M. (2017). Xsl transformations (xslt) version 3.0. [online]. [Accessed 2023-08-13].

Khare, R. (2006). Microformats: the next (small) thing on the semantic web? *IEEE Internet Computing*, 10(1):68–75.

Lin, D., Crabtree, J., Dillo, I., Downs, R., Edmunds, R., Giaretta, D., Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D., Stockhause, M., and Westbrook, J. (2020). The trust principles for digital repositories. *Scientific Data*, 7.

Martínková, J. and Suchánek, M. (2024). Semantically annotated data management plans. [Accessed 2024-04-03].

Miksa, T., Walk, P., Neish, P., Oblasser, S., Holland, M., Renner, T., Jacquemot-Perbal, M.-C., Cardoso, J., Kvamme, T., Praetzellis, M., et al. (2021). Application Profile for Machine-Actionable Data Management Plans.

National Institutes of Health (2023). Data management & sharing plan. [online]. [Accessed 2023-08-13].

Open Knowledge (2015). Machine-readable. [Accessed 2023-07-17].

OpenAI (2022). ChatGPT (version 3.5). [Accessed 2023-12-15].

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*, 18:59.

RDFa Working Group (2013). RDF in Attributes (RDFa). [Accessed 2023-07-17].

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Science Europe (2021). Practical guide to the international alignment of research data management-extended edition.

Smale, N., Unsworth, K., Denyer, G., and Barr, D. (2018). The history, advocacy and efficacy of data management plans. *bioRxiv*.

W3C (2012). W3c xml schema definition language (xsd) 1.1. [online]. [Accessed 2023-08-15].

W3C (2017). The extensible stylesheet language family (xsl). [online]. [Accessed 2023-08-15].

Web Hypertext Application Technology Working Group (2023). Html - living standard. [online]. [Accessed 2023-12-15].

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2020). Transformers: State-of-the-art natural language processing. [Accessed 2023-12-15].