Automated Semantic Annotation of Data Management Plans: A Systematic Review

REVIEW

JANA MARTÍNKOVÁ
MAREK SUCHÁNEK
PETR KROHA

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

ABSTRACT

Semantic annotation has emerged as a key technique for transforming human-readable data into machine-actionable formats. It corresponds with the growing emphasis on data reusability and research reproducibility. This paper examines tools for semantic annotation using ontologies and controlled vocabularies, with a focus on their application in data management planning. A systematic review identified 34 relevant tools, which show potential for adaptation to the data management plan (DMP) domain. While these tools meet many requirements, they do not fully address all DMP-specific needs. The paper provides an overview of current tools and suggests directions for future research to adapt them for DMP use.

CORRESPONDING AUTHOR:
Jana Martínková
Faculty of Information Technology, CTU in Prague, Thákurova, Prague, 160 00, Czech Republic
jana.martinkova@cvut.cz

KEYWORDS:
semantic annotation; data management plans; ontology; tool; systematic literature review

TO CITE THIS ARTICLE:
Martínková, J., Suchánek, M. and Kroha, P. 2025 Automated Semantic Annotation of Data Management Plans: A Systematic Review. *Data Science Journal*, 24: 16, pp. 1–14. DOI: https://doi.org/10.5334/dsj-2025-016

DOI: 10.5334/dsj-2025-016

1 INTRODUCTION

When filling out forms and questionnaires in connection with applications for research grants, it is necessary for the applicant to correctly understand the meaning of the words used to request information, i.e., doing it so that the answers are semantically consistent with the intention of the author of the questionnaire. Because of the ambiguity of natural language, this is a non-trivial problem. It is necessary to unify the semantic meaning of the words between the author of the questionnaire and the applicant. Simplificated, semantic annotations can be denoted as explanations. These annotations are not only intended to clarify meaning for humans but also to make the information machine-actionable, enabling automated systems to process, interpret, and utilize the data correctly. The problem is how to generate them automatically.

In recent years, data management planning has gained significant importance, largely due to the requirements set by funding agencies and research institutions. Data management planning is a comprehensive process that spans the entire data life cycle, applying key data management practices to ensure accurate data collection, secure storage, proper handling, and potential reuse beyond the primary project (Smale *et al.*, 2018).

A key element of data management planning is the development and ongoing maintenance of a document known as the data management plan (DMP). This document contains detailed responses to questions about data management practices that are required by funding agencies or institutions. A completed data management plan (DMP) contains valuable details about the data used and/or collected in the project, enabling research reproducibility and facilitating data reuse by other researchers. This reuse helps avoid duplicating efforts and maximizes data value (DataCite, 2021).

While there are existing tools to aid in creating DMPs, the burden of manually inputting information remains a significant challenge. Researchers, often with the support of data stewards, invest considerable time and effort in completing a DMP, which typically results in a document designed primarily for human readers—most often for the benefit of funding agencies. Although the idea of the DMP is sound, its human-readable format limits the usability of the rich information it contains.

There are emerging approaches to creating machine-actionable Data Management Plans (maDMPs). Still, the complexity of the data described in DMPs often makes it difficult to represent fully in structured formats. As a result, even in the Research Data Alliance (RDA) standard for maDMPs (Miksa *et al.*, 2021), there are sections of free text that cannot be fully understood or interpreted by machines.

This limitation leads to the exploration of automated semantic annotation as a potential solution, where human-readable text is transformed to include interpretation that machines can directly process and utilize. This study aims to provide an overview of tools for semantic annotation using ontologies, offering insights into possible techniques that could improve the machine-actionability of DMPs.

To better illustrate the challenges associated with automated semantic annotation in the DMP domain, we present two illustrative examples (see Code Example 1). These examples showcase the typical structure of a DMP, where bold sections represent the questions or guidelines, and the accompanying text reflects the responses provided by the researcher or data steward. Both examples are taken from existing DMP,[1] which adheres to the Horizon Europe Template (European Commission, 2022).

Principal Investigator: Thomas Pölzler

What is the expected size of the data that you intend to generate or reuse?

I expect that the total amount of data (mostly .xls, .csv, .pdf, and .sav files) will be < 100 MB, including all project outputs such as journal articles.

Code Example 1 Illustrative examples of DMP.

1 Pölzler, T. (2023). Data Management Plan: Making Morality Impartial: An Experimental Investigation of the Veil of Ignorance. Available at https://dmponline.dcc.ac.uk/public_plans.

2 RELATED WORK

The primary objective of the DMP Common Standards Working Group (DCS WG),[2] operating under the RDA, is to establish well-defined processes for research data management, develop a robust infrastructure, and, most importantly, create a universally accepted standard for representing DMP information (Miksa *et al.*, 2021). As part of these efforts, the DCS WG has introduced a JSON serialization of their application profile, providing a practical framework for implementation.

However, significant challenges remain with the DCS WG's application profile, as highlighted by Cardoso *et al.*, (2022). One major limitation is the absence of explicit linkages to existing ontologies, which impedes semantic integration and interoperability. Furthermore, the DMP Common Standards (DCS) profile covers only the most essential aspects of DMPs, omitting critical details such as the provenance of reused or generated data, project objectives, data access embargoes, and access protocols. Many of the existing fields are also free-text, allowing for the input of diverse and inconsistent information, which renders them difficult to interpret or process by machines.

In response to these limitations, Cardoso *et al.* (2022) developed the DMP Common Standard Ontology (DCSO), which builds upon the DCS standard specification. This development adds semantic meaning and structure, providing a step forward in connecting DMPs with ontologies. However, despite this advancement, the DCSO still falls short of fully encompassing the comprehensive content required for DMPs.

In our previous work (Martínková and Suchánek, 2023), we examined and mapped ontologies and controlled vocabularies relevant to the DMP domain, including the DCSO (Cardoso *et al.*, 2022), identifying inconsistencies and overlaps in terms and concepts. This analysis revealed that inconsistencies in terminology are also prevalent in existing DMP templates, further complicating semantic interoperability. To address these issues, we developed an OntoUML conceptual model (Martínková *et al.*, 2024) that provides a comprehensive representation of DMP content with an emphasis on accurate terminology, clear relationships, and full content coverage. We are currently developing a properly defined ontology based on this model.

In parallel, we conducted a study (Martínková and Suchánek, 2024) to explore approaches for capturing DMPs in a dual format that is both machine-actionable and human-readable. This study found that a hybrid approach combining manual methods and automated techniques (e.g., leveraging artificial intelligence (AI)) shows significant promise.

To illustrate these challenges and opportunities, we present two simple examples (Code Example 2) demonstrating the approach that combines machine-actionable and human-readable formats using annotations in Resource Description Framework in Attributes (RDFa) (RDFa Working Group, 2013). It is important to note that these examples are not exhaustive; they are modified for illustrative purposes to highlight the types of information that become relevant when dealing with DMPs.

In the second example, for instance, the responsew includes additional details, such as the data formats used. However, the DMP template often contains separate sections dedicated to format information. This illustrates a crucial point: while the response mentions formats, the focus of this specific question is on the volume of data. Such distinctions are critical for maintaining consistency and avoiding redundancy. Colors are used in the examples to distinguish different elements for illustration purposes.

Principal Investigator: Thomas Pölzler

What is the expected size of the data that you intend to generate or re-use? I expect that the total amount of data (mostly .xls, .csv, .pdf and .sav files) will be < 100 MB , including all project outputs such as journal articles.

Code Example 2 Illustrative examples of DMP with annotations.

2 https://www.rd-alliance.org/groups/dmp-common-standards-wg.

To further investigate these possibilities, we undertook this review to systematically explore existing tools and methodologies for semantic annotation and their applicability to DMPs.

Notably, despite the increasing attention to maDMPs, no existing review, to our knowledge, focuses specifically on semantic annotation tools or methodologies in the DMP domain. This gap in the literature underscores the novelty and necessity of our research, which aims to bridge the intersection of semantic annotation and data management planning.

3 METHODOLOGY

In this work, two well-known methods for literature review are used: the Systematic Literature Review (SLR) method (Lame, 2019; Nightingale, 2009) for defining the initial set of studies and the snowballing method (Wohlin, 2014) as a search approach.

The SLR method is employed to identify an initial set of studies addressing our research question. Initially developed in the health sciences, this method has been widely adopted in other disciplines due to its strengths, including a structured approach to answering key research questions and enhancing the understanding and monitoring of research practices across diverse fields (Lame, 2019). The concept of evidence-based software engineering was introduced by Kitchenham *et al*. (2004), promoting systematic reviews to improve the rigor and relevance of software engineering research. This evidence-based approach has also been adapted in related fields, such as information systems research, as demonstrated by the work of Webster and Watson (2002), who advocated for similar structured methods that enhance literature reviews and the development of theoretical foundations.

By adhering to the SLR process, the literature is reviewed systematically, minimizing bias by preventing authors from prioritizing well-known studies or those that align with their personal opinions or prior research outcomes. To ensure that all relevant studies addressing the research question are captured, the aims and objectives must be clearly defined, along with inclusion and exclusion criteria that guide both the review process and the search strategy (Nightingale, 2009). These criteria are critical to refining the scope of the review and ensuring that the studies selected are directly relevant to the research objectives.

The snowballing method (Wohlin, 2014) complements the SLR method by serving as an additional search strategy, as illustrated in Figure 1. In this approach, relevant studies are identified by examining the reference lists (backward snowballing) of the initial set of studies, such as those obtained from the SLR, or by analyzing where they have been cited (forward snowballing). This iterative process continues until no new relevant studies are found in either direction, ensuring comprehensive coverage of the literature related to our research question.

An important recommendation from Wohlin (2014) is to include highly cited papers in the initial set of studies, as these often represent significant contributions to the field. Following this guideline, we included key studies such as Miksa *et al*. (2021) and Cardoso *et al*. (2022). These papers are influential in the area of machine-actionable DMPs, as outlined in Section 2. Additionally, we planned to use a search engine to identify any additional tools, as not all relevant tools necessarily have publications. This ensures we do not overlook any valuable resources.

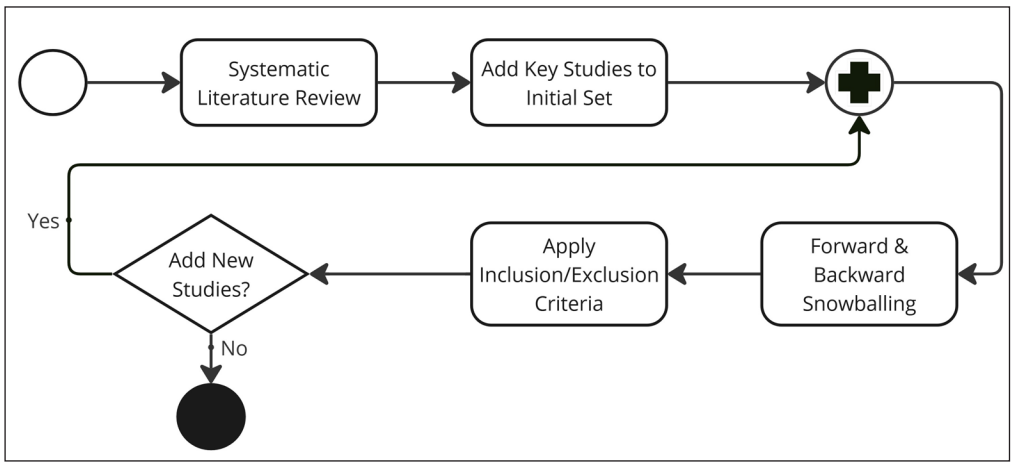

Figure 1 The methodology diagram (own adaptation according to Wohlin (2014)).

4 LITERATURE REVIEW

In our literature review, we followed the methodology outlined in Section 3, employing the SLR method to identify the initial set of studies. We began by defining the research question, which informed the formulation of the search query used in the SLR process to obtain a starting set of studies relevant to our research question. Additionally, we established filters and evaluation criteria to refine the search results and ensure the relevance and quality of the studies selected.

After conducting the SLR and obtaining the initial set of studies for final analysis, we proceeded with the review by applying the snowballing method iteratively, using both forward and backward snowballing to identify additional relevant studies. No new results were obtained after the fifth iteration, resulting in a total of 34 tools.

4.1 RESEARCH QUESTION

The objective of this study is to explore possible approaches to the automated semantic annotation of DMPs to produce DMP documents that retain human readability while gaining machine-actionability through the use of ontologies and/or controlled vocabularies in the annotation process. This leads to the formulation of the following research question:

> *What are the existing approaches to the automated semantic annotation of DMPs using ontological terms?*

This research question aims to provide an overview of current approaches, primarily in the context of DMPs. However, approaches from other fields may also be relevant. Therefore, we will include studies that explore more general approaches to automated semantic text annotation, even if they do not specifically address DMPs or use ontologies. Expanding the scope in this way allows for a broader understanding of potential methodologies that could be applied to the semantic annotation of DMPs. The complete search query related to this research question is provided in Appendix A.

4.2 SEARCH AND FILTERS

Following the approach recommended in Nightingale (2009), we aimed to capture a broad range of relevant studies by focusing on sensitivity rather than specificity. This approach minimizes bias by incorporating various synonyms and related terms.

For our research, we used Google Scholar as an academic database because it offers broad access to publications across various disciplines and helps to avoid bias toward specific publishers, which is also recommended by Nightingale (2009). We performed the search in October and November 2024. We filtered the results and included only papers in English and published from the year 2015 onward.

The initial query returned 84 results. As anticipated, none of these focused on the DMP domain, so we broadened our search to include other related fields, making the research more general. After removing duplicates and filtering for relevance, only 16 results were directly aligned with our focus. These primarily consisted of reviews discussing various tools and methodologies across multiple domains. To further expand our scope, we applied both forward and backward snowballing techniques, allowing us to identify 34 tools.

To identify additional results that might not have associated publications, we conducted a Google search using the same initial query. We chose Google due to its dominant role in the search engine market, with an estimated 90% share.[3] The search, conducted in March 2025, returned 115 results, approximately 100 of which were new entries not identified in the earlier Google Scholar search. However, none of these newly discovered results were directly relevant to our focus or met our filtering criteria.

Semantic annotation has gained significant prominence in recent years, with a wide range of methods and tools emerging to address various challenges. Due to the practical and application-oriented nature of this field, we chose to focus this review specifically on the tools available, as they provide direct solutions to real-world problems. While these tools are crucial for enabling semantic annotation, we recognize that the underlying approaches are equally

3 https://gs.statcounter.com/search-engine-market-share.

important for advancing the field. Therefore, we will dedicate a subsequent paper to exploring these methods in greater detail. This exploration will examine their theoretical foundations, methodologies, and their potential applicability across different domains.

4.3 EVALUATION CRITERIA

The following criteria were defined to systematically evaluate the applicability of each identified tool for the semantic annotation of DMPs. Given the multidimensional nature of these criteria, numerical evaluation was inappropriate. Instead, we used three categories (or two in some cases) to assess the extent to which each tool meets the criteria.

Use of Ontologies or Controlled Vocabularies. This criterion evaluates whether the tool employs ontologies, controlled vocabularies, or taxonomies for annotation. Additionally, it evaluates whether the tool supports the use of custom ontologies or restricts it to predefined ones.

- ●: The tool uses ontologies or controlled vocabularies and allows customization with user-defined ontologies.
- ◐: The tool uses ontologies or controlled vocabularies, but customization is either unsupported or undocumented.
- ○: The tool does not utilize ontologies or controlled vocabularies.

Automation. This criterion assesses the degree of automation provided by the tool in performing semantic annotation tasks.

- ●: The tool is fully automated, requiring no manual intervention.
- ◐: The tool is semi-automated, combining automation with manual adjustments.
- ○: The tool relies primarily on manual processes.

Input. This criterion evaluates the compatibility of the tool with plain text as an input format, which is the typical form in which DMPs are presented.

- ●: The tool is fully compliant with plain text input.
- ○: The tool is incompatible with plain text input or unclear about the required input format.

Output. This criterion evaluates the formats of the output produced by the tool, with a particular focus on ensuring they are machine-readable.

- ●: Supports machine-readable formats such as Resource Description Framework (RDF), JavaScript Object Notation for Linked Data (JSON-LD), or Extensible Markup Language (XML).
- ◐: Produces basic formats like plain text or CSV, which may lack advanced interoperability.
- ○: Provides unique or configurable output formats, or the output format is unspecified.

Maturity and Applications. This criterion evaluates the maturity of the tool, along with its real-world applications and use cases.

- ●: The tool is actively used in production environments.
- ◐: The tool is a research prototype with limited real-world applications.
- ○: The tool is either no longer maintained or supported, or there is no available information about its current usage.

Licensing and Accessibility. This criterion defines the usage rights and accessibility of the tool.

- ●: The tool is open-source and freely available for modification and use.
- ◐: The tool is free for academic or personal use but has restrictions for commercial use.
- ○: The tool requires a license or subscription for access, or the licensing status is unspecified or unclear.

Evaluation Undertaken. This criterion assesses whether the tool or approach has been empirically evaluated using metrics, datasets, benchmarks, or comparisons with other approaches or tools.

- ●: Comprehensive evaluation is provided, including relevant metrics, benchmarks, and comparisons.
- ◐: Some evaluation is provided, with limited metrics, datasets, or benchmarks mentioned.
- ○: No empirical evaluation or performance metrics are mentioned.

4.4 RESULTS

We evaluated, based on our criteria, all tools that claimed to apply semantic annotation, and the results are summarized in Table 1. The tools in this table are listed in alphabetical order.

Table 1 Semantic annotation tools and their compliance with the criteria.

TOOL	USE OF ONTOLOGY	AUTOMATION	INPUT	OUTPUT	MATURITY AND APPLICATIONS	LICENSING AND ACCESSIBILITY	EVALUATION UNDERTAKEN
ABNER (Settles, 2005)	○	●	●	●	●	●	●
BeCAS (Nunes *et al.*, 2013)	◐	●	●	●	◐	●	○
Bio-YODIE (Gorrell *et al.*, 2018)	◐	●	●	◐	●	●	●
cTAKES (Savova *et al.*, 2010)	●	●	●	○	◐	●	●
Cerno (Kiyavitskaya *et al.*, 2009)	◐	◐	●	●	○	○	●
ChemSpot (Rocktäschel *et al.*, 2012)	◐	●	●	○	◐	◐	●
ConceptMapper (Tanenblatt *et al.*, 2010)	●	●	○	●	●	●	●
CONANN (Reeve and Han, 2007)	◐	●	●	◐	○	○	●
ContracT (Soavi *et al*, 2020)	◐	◐	●	○	○	○	◐
EDGAR (Rindflesch *et al.*, 1999)	◐	●	●	●	◐	○	○
GENIES (Friedman *et al.*, 2001)	◐	◐	●	◐	◐	○	◐
Huang et al. (Huang *et al.*, 2006)	◐	●	●	●	◐	○	●
KIM (Popov *et al.*, 2003)	●	●	●	●	◐	○	◐
Marvin (Milosevic, 2016)	●	●	●	●	◐	●	○
MedCAT (Kraljevic *et al.*, 2021)	●	●	●	○	●	●	●
MetaMap (Aronson, 2001)	◐	●	●	○	●	○	●
NBCO Annotator (Jonquet *et al.*, 2009)	◐	●	○	●	◐	●	○
Neural Concept Recognizer (Arbabi *et al.*, 2019)	●	●	●	◐	●	●	●
NOBLE Coder (Tseytlin *et al.*, 2016)	●	●	○	◐	◐	●	●
OnTeA (Laclavik *et al.*, 2006)	◐	●	●	○	◐	○	●
OntoBlog (Shakya *et al.*, 2007)	◐	◐	●	○	◐	○	◐
OPTIMA (Vlachidis and Tudhope, 2016)	◐	●	●	●	◐	○	●
OSCAR4 (Jessop *et al.*, 2011)	●	●	○	◐	●	◐	●
PASTA (Gaizauskas *et al.*, 2003)	◐	●	●	◐	◐	●	○
PANKOW (Cimiano *et al.*, 2004)	●	●	●	●	○	○	◐
RysannMD (Cuzzola *et al.*, 2017)	◐	●	●	◐	◐	○	●
SemTag (Dill *et al.*, 2003)	◐	●	○	●	◐	○	●
SiGEG (Haghgoo *et al.*, 2022)	●	●	●	●	◐	◐	●
SnoMedTagger (Hina *et al.*, 2013)	◐	◐	●	◐	◐	○	●
TaggerOne (Leaman and Lu, 2016)	●	●	●	○	◐	●	●
Textpresso (Müller *et al.*, 2004)	◐	●	●	●	●	●	◐
Verdant (McKain *et al.*, 2017)	○	●	○	○	●	●	◐
Whatizit (Rebholz-Schuhmann *et al.*, 2008)	◐	●	●	○	◐	◐	○
XONTO (Oro and Ruffolo, 2008)	◐	◐	●	◐	○	○	○

The evaluation reveals that the majority of tools align with our research question and incorporate ontologies in their semantic annotation processes. However, if we were to use one of these tools directly for the DMP domain, ontology customization would be essential. Unfortunately, this feature is supported by less than half of the tools evaluated.

Most tools are fully automated, enabling efficient processing with minimal user intervention, though six are semi-automatic and require some level of human involvement. Regarding input compatibility, many tools accept plain text, making them versatile for unstructured data. However, only a few fully support outputs in machine-actionable formats, limiting their interoperability and utility for our case.

A notable observation is that most tools are research prototypes and not widely adopted in real-world scenarios. This is often linked to the lack of detailed information about their licensing and accessibility. While some tools provide comprehensive evaluations, others lack sufficient details about their performance, leaving questions about their reliability and scalability unanswered.

5 DISCUSSION

Semantic annotation tools vary significantly in terms of features, usability, and methodologies, reflecting the diverse needs they address. Highly rated tools such as ConceptMapper (Tanenblatt *et al.*, 2010), ABNER (Settles, 2005), MedCAT (Kraljevic *et al.*, 2021), TextPresso (Müller *et al.*, 2004), Bio-YODIE (Gorrell *et al.*, 2018), and Neural Concept Recognizer (Arbabi *et al.*, 2019) demonstrate a balance of automation, open-source availability, maturity, and empirical evaluation, making them strong candidates for our applications. However, gaps in ontology integration, input/output format flexibility, and adaptability to specific domains highlight areas that do not fully align with our requirements.

ConceptMapper (Tanenblatt *et al.*, 2010) is a dictionary-based named entity recognition tool that links biomedical entities in clinical text to medical ontologies. Its simplicity and reliance on predefined dictionaries make it effective for tasks where vocabulary is well-defined but less adaptable to novel data.

ABNER (Settles, 2005), on the other hand, employs a machine-learning approach using Conditional Random Fields to identify entities such as proteins or DNA. While effective for recognition, its lack of ontology integration limits its utility for tasks requiring semantic linking or reasoning.

MedCAT (Kraljevic *et al.*, 2021) is a machine-learning tool designed for electronic health records. It excels at recognizing and linking biomedical entities to customizable medical ontologies, making it particularly suitable for domains with extensive, curated vocabularies.

TextPresso (Müller *et al.*, 2004) is an ontology-based text mining tool designed to annotate and search biological literature. It annotates terms in articles and abstracts using a predefined ontology comprising 33 categories that represent various biological concepts and relationships.

Bio-YODIE (Gorrell *et al.*, 2018) is a named entity linking system for biomedical text that identifies entities and maps them to medical ontologies. It employs predefined domain-specific rules and patterns for entity annotation, which makes adaptation to other domains challenging.

The Neural Concept Recognizer (Arbabi *et al.*, 2019) employs Convolutional Neural Networks (CNNs) and ontology embedding to identify and map text phrases to biomedical ontologies, including previously unseen synonyms. By integrating pre-trained word embeddings with hierarchical ontology structures, it achieves improved accuracy and effectively handles the complexity of the biomedical domain.

Despite these strengths, the tools reviewed reveal methodological trade-offs and limitations that are crucial to understanding their applicability. For instance, rule-based methods excel in precision but require significant manual effort to define rules, limiting scalability. Machine learning methods, while adaptable and powerful, often demand large annotated datasets, which may not be available for all domains, such as in the case of DMP. Similarly, ontology-based approaches enable semantic consistency and reasoning but are often underutilized, as many tools treat ontologies merely as dictionaries rather than leveraging their full semantic depth. These limitations are further highlighted by applying our two examples to a selection of the tools. The purpose of this exercise was not only to explore their core functionality but also to see how effectively they can be adapted to meet the unique requirements of our domain. While these tools were primarily

designed for other domains, we sought to explore their potential for generalization and identify opportunities for improvement or modification to better align with our needs.

ConceptMapper (Tanenblatt *et al.*, 2010) successfully identified the name while mistakenly identifying the role of 'Principal Investigator' as a name as well, as shown in Figure 2. MedCAT (Kraljevic *et al.*, 2021) was able to correctly recognize the data in the later example as 'data', as illustrated in Figure 3, but nothing else. ABNER (Settles, 2005), however, did not identify any entities, and the Neural Concept Recognizer (Arbabi *et al.*, 2019) similarly failed to recognize anything. We found available TextPresso (Müller *et al.*, 2004) that only provided a search tool for previously annotated texts but did not offer annotation capabilities, which makes it unsuitable for our use case. Bio-YODIE (Gorrell *et al.*, 2018) was not executable on our machines, possibly due to lack of maintenance and updates; its last code commit was made 6 years ago.[4]

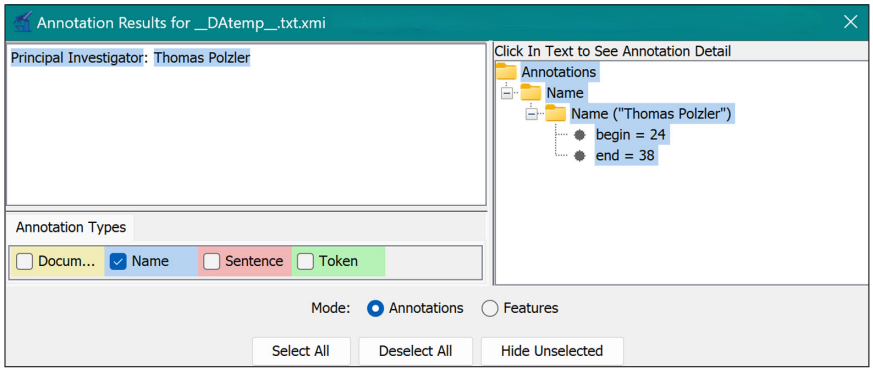

Figure 2 ConceptMapper (Tanenblatt *et al.*, 2010) successfully identified the name but unsuccessfully identified the role, which is also highlighted in blue.

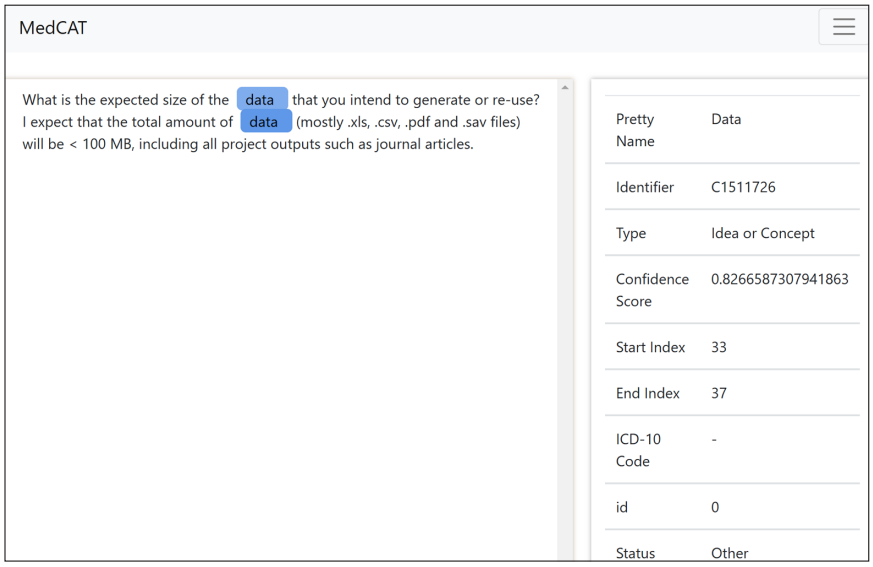

Figure 3 MedCAT (Kraljevic *et al.*, 2021) successfully identified the data.

These observations underscore the need to explore methodologies to identify common approaches, their limitations, and how they align with our research needs. Many tools employ multiple methodologies across different steps in their workflows, adapting to the challenges of specific tasks. To provide a clearer understanding, we mapped the reviewed tools to their underlying methodologies (Table 2).

The classifications in Table 2 are based on the descriptions provided in the respective articles and our interpretation of the methodologies. Below is a summary of the methodologies observed:

- *Rule-based* methods rely on predefined rules, dictionaries, or regular expressions crafted by experts. These methods are highly precise but require significant manual effort to adapt to new domains or tasks.
- *Machine learning-based* methods train statistical or deep learning models on annotated datasets. While these methods offer adaptability and high performance, they require large amounts of annotated data, which is often a limitation.

4 https://github.com/GateNLP/Bio-YODIE.

- *Ontology-based* methods leverage structured knowledge representations to match text entities with ontology terms. These methods ensure semantic consistency but are often underutilized (as indicated by the ◐ in the table), with many tools using ontologies primarily for term lookups.
- *Pattern-based* methods focus on linguistic or structural patterns in text, either manually defined or learned automatically.
- *Dictionary-based* methods match text against predefined dictionaries of terms and phrases. While straightforward and easy to implement, they are limited by the scope and quality of the dictionaries.

This exploration highlights the diverse methodologies employed in semantic annotation. For tasks that lack large annotated datasets, rule-based and dictionary-based approaches may provide more reliable results, while machine learning methods excel in contexts with a lot of training data.

Table 2 Semantic annotation tools and their methodologies.

TOOL	RULE-BASED	MACHINE LEARNING	ONTOLOGY-BASED	PATTERN-BASED	DICTIONARY-BASED
ABNER (Settles, 2005)	○	●	○	○	○
BeCAS (Nunes *et al.*, 2013)	○	○	◐	○	●
Bio-YODIE (Gorrell *et al.*, 2018)	●	◐	◐	○	●
Cerno (Kiyavitskaya *et al.*, 2009)	○	○	○	●	○
ChemSpot (Rocktäschel *et al.*, 2012)	○	●	○	○	●
ConceptMapper (Tanenblatt *et al.*, 2010)	○	○	○	○	●
CONANN (Reeve and Han, 2007)	●	○	◐	●	●
ContracT (Soavi *et al.*, 2020)	●	○	○	○	●
cTAKES (Savova *et al.*, 2010)	●	●	◐	○	●
EDGAR (Rindflesch *et al.*, 1999)	○	○	◐	○	●
GENIES (Friedman *et al.*, 2001)	●	○	●	●	○
Hunag et all (Huang *et al.*, 2006)	●	○	●	●	○
KIM (Popov *et al.*, 2003)	○	○	●	○	○
Marvin (Milosevic, 2016)	○	○	◐	○	●
MedCAT (Kraljevic *et al.*, 2021)	○	●	◐	○	○
MetaMap (Aronson, 2001)	○	○	◐	○	●
NBCO annotator (Jonquet *et al.*, 2009)	○	○	●	○	●
Neural Concept Recognizer (Arbabi *et al.*, 2019)	○	●	●	○	○
NOBLE Coder (Tseytlin *et al.*, 2016)	●	○	◐	○	●
OnteA (Laclavık *et al.*, 2006)	○	○	◐	●	○
OntoBlog (Shakya *et al.*, 2007)	●	○	◐	○	●
OPTIMA (Vlachidis and Tudhope, 2016)	●	○	◐	○	○
OSCAR4 (Jessop *et al.*, 2011)	●	●	○	●	●
PASTA (Gaizauskas *et al.*, 2003)	●	○	●	○	○
PANKOW (Cimiano *et al.*, 2004)	○	○	○	●	○
rysannMD (Cuzzola *et al.*, 2017)	●	○	○	○	●
SemTag (Dill *et al.*, 2003)	○	○	○	●	○
SiGEG (Haghgoo *et al.*, 2022)	○	●	◐	●	○
SnoMedTagger (Hina *et al.*, 2013)	●	○	◐	○	●
TaggerOne (Leaman and Lu, 2016)	○	●	◐	○	○
Textpresso (Müller *et al.*, 2004)	○	○	◐	●	●
Verdant (McKain *et al.*, 2017)	●	○	○	○	●
Whatizit (Rebholz-Schuhmann *et al.*, 2008)	○	○	◐	○	●
XONTO (Oro and Ruffolo, 2008)	●	○	◐	●	●

6 CONCLUSION

We conducted a systematic literature review to assess existing approaches for the automated semantic annotation of DMP documents using ontologies. Although our methodology followed a structured and rigorous process, the study's limitations primarily arise from the challenges in identifying all relevant literature. Despite our systematic efforts, it is possible that some relevant publications were not captured, for instance, due to incomplete indexing, insufficient metadata, or unconventional keyword usage in the original sources.

This study examined various tools designed for semantic annotation using ontologies. The high volume of research in this area underscores its importance in transforming human-readable data into machine-actionable formats. However, we identified a notable gap in tools tailored to the DMP domain. Consequently, we expanded our scope to explore tools from a wide range of fields, including biomedicine, where such tools have seen significant advancements.

Our evaluation highlighted a range of tools employing diverse methodologies for semantic annotation. Among these, ConceptMapper (Tanenblatt *et al.*, 2010), ABNER (Settles, 2005), MedCAT (Kraljevic *et al.*, 2021), TextPresso (Müller *et al.*, 2004), Neural Concept Recognizer (Arbabi *et al.*, 2019), and Bio-YODIE (Gorrell *et al.*, 2018) emerged as the most relevant to our needs. While these tools meet certain requirements, they do not fully address the unique needs of our approach in the DMP domain. Building on these findings, we intend to explore the applicability of these tools for the DMP domain despite their limitations. At the same time, we acknowledge the need for a novel solution and propose to leverage the knowledge and methodologies of these tools to develop an approach specifically tailored to the requirements of DMPs. To facilitate this, we have mapped the reviewed tools to their underlying methodologies, offering an overview of the techniques commonly used in automated semantic annotation tools.

Future work will delve into the methods that support these tools and explore their potential to enhance data management practices in research. Building on the insights gained, we will propose an approach to align more closely with best practices and advance data management in research.

APPENDIX A Search query

("semantic annotation" OR "semantic labeling" OR "semantic enrichment") AND ("text mining" OR "natural language processing" OR "NLP" OR "machine learning" OR "artificial intelligence") AND ("ontology" OR "ontologies" OR "ontology-based") AND ("document annotation" OR "text annotation" OR "automated annotation" OR "information extraction") AND ("Data Management Plans" OR "structured documents" OR "metadata" OR "research plans" OR "data documentation").

FUNDING INFORMATION

This work was supported by the Czech Technical University in Prague grant No. SGS23/206/OHK3/3T/18, and the Ministry of Education, Youth and Sports grant No. LM2023055.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Jana Martínková orcid.org/0000-0001-8575-6533
Faculty of Information Technology, CTU in Prague, Thákurova, Prague, 160 00, Czech Republic

Marek Suchánek orcid.org/0000-0001-7525-9218
Faculty of Information Technology, CTU in Prague, Thákurova, Prague, 160 00, Czech Republic

Petr Kroha orcid.org/0000-0002-1658-3736
Faculty of Information Technology, CTU in Prague, Thákurova, Prague, 160 00, Czech Republic

REFERENCES

Arbabi, A., Adams, D.R., Fidler, S. and **Brudno, M.** (2019) 'Identifying clinical terms in free-text notes using ontology-guided machine learning', in L.J. Cowen (ed.) *Research in computational molecular biology*. Cham: Springer International Publishing, pp. 19–34. Available at: https://doi.org/10.1007/978-3-030-17083-7_2

Aronson, A.R. (2001) 'Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program', in *Proceedings of the AMIA Symposium*, Washington, DC, 3–7 November. American Medical Informatics Association, pp. 17–21.

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), p. 21. Available at: https://doi.org/10.1186/s13326-022-00274-4

Cimiano, P., Handschuh, S. and **Staab, S.** (2004) 'Towards the self-annotating web', *The 13th international conference on world wide web*. New York, NY, 17–20 May. New York, NY: Association for Computing Machinery, pp. 462–471. Available at: https://doi.org/10.1145/988672.988735

Cuzzola, J., Jovanović, J. and **Bagheri, E.** (2017) 'RysannMD: A biomedical semantic annotator balancing speed and accuracy', *Journal of Biomedical Informatics*, 71, pp. 91–109. Available at: https://doi.org/10.1016/j.jbi.2017.05.016

DataCite. (2021) *Introduction to machine actionable dmps (madmps)*. Available at: https://support.datacite.org/docs/introduction-to-machine-actionable-dmps-madmps.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A. and **Zien, J.Y.** (2003) 'SemTag and seeker: Bootstrapping the semantic web via automated semantic annotation', *The 12th international conference on world wide web*. Budapest, Hungary, 20–24 May. New York, NY: Association for Computing Machinery, pp. 178–186. Available at: https://doi.org/10.1145/775152.775178

European Commission. (2022) *Horizon europe data management plan template*. Available at: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/report/data-management-plan˙he˙en.docx.

Friedman, C., Kra, P., Yu, H., Krauthammer, M. and **Rzhetsky, A.** (2001) 'GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles', *Bioinformatics*, 17, pp. S74–S82. Available at: https://doi.org/10.1093/bioinformatics/17.suppl_1.S74

Gaizauskas, R., Demetriou, G., Artymiuk, P.J. and **Willett, P.** (2003) 'Protein structures and information extraction from biological texts: The PASTA System', *Bioinformatics*, 19(1), pp. 135–143. Available at: https://doi.org/10.1093/bioinformatics/19.1.135

Gorrell, G., Song, X. and **Roberts, A.** (2018) 'Bio-YODIE: A named entity linking system for biomedical text', *arXiv*, arXiv:1811.04860. Available at: https://doi.org/10.48550/arXiv.1811.04860

Haghgoo, M., Nazary, A.N.A. and **Monti, A.** (2022) 'SiSEG-auto semantic annotation service to integrate smart energy data', *Energies*, 15(4), p. 1428. Available at: https://doi.org/10.3390/en15041428

Hina, S., Atwell, E. and **Johnson, O.A.** (2013) 'SnoMedTagger: A semantic tagger for medical narratives', *International Journal of Computational Linguistics*, 4(2), pp. 81–99.

Huang, Y.T., Yeh, H.Y., Cheng, S.W., Tu, C.C., Kuo, C.L. and **Soo, V.W.** (2006) 'Automatic extraction of information about the molecular interactions in biological pathways from texts based on ontology and semantic processing', *2006 IEEE International Conference on Systems, Man, and Cybernetics*. Taipei, Taiwan, 8–11 October. Institute of Electrical and Electronics Engineers, pp. 3679–3684. Available at: https://doi.org/10.1109/ICSMC.2006.384701

Jessop, D.M., Adams, S.E., Willighagen, E.L., Hawizy, L. and **Murray-Rust, P.** (2011) 'OSCAR4: A flexible architecture for chemical text-mining', *Journal of Cheminformatics*, 3(1), p. 41. Available at: https://doi.org/10.1186/1758-2946-3-41

Jonquet, C., Shah, N.H. and **Musen, M.A.** (2009) 'The open biomedical annotator', *Summit on Translational Bioinformatics*, 2009, pp. 56–60.

Kitchenham, B.A., Dyba, T. and **Jorgensen, M.** (2004) 'Evidence-based software engineering. In: Proceedings', *26th international conference on software engineering*. Edinburgh, UK, 28 May. Institute of Electrical and Electronics Engineers, pp. 273–281.

Kiyavitskaya, N., Zeni, N., Cordy, J.R., Mich, L. and **Mylopoulos, J.** (2009) 'Cerno: Light-weight tool support for semantic annotation of textual documents', *Data & Knowledge Engineering*, 68(12), pp. 1470–1492. Available at: https://doi.org/10.1016/j.datak.2009.07.012

Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., Mascio, A., Zhu, L., Folarin, A.A., Roberts, A., Bendayan, R., Richardson, M.P., Stewart, R., Shah, A.D., Wong, W.K., Ibrahim, Z., Teo, J.T. and **Dobson, R.J.B.** (2021) 'Multi-domain clinical natural language processing with MedCAT: the Medical Concept Annotation Toolkit', arXiv:2010.01165. Available at: https://doi.org/10.48550/arXiv.2010.01165

Laclavık, M., Šeleng, M. and **Babık, M.** (2006) 'Ontea: Semi-automatic ontology based text annotation method', *Tools for Acquisition, Organisation and Presenting of Information and Knowledge*, pp. 49–63.

Lame, G. (2019) 'Systematic literature reviews: An introduction', *Proceedings of the Design Society: Inernational Conference on Engineering Design*, 1(1), pp. 1633–1642. Available at: https://doi.org/10.1017/dsi.2019.169

Leaman, R. and **Lu, Z.** (2016) 'TaggerOne: Joint named entity recognition and normalization with semi-Markov Models', *Bioinformatics*, 32(18), pp. 2839–2846. Available at: https://doi.org/10.1093/bioinformatics/btw343

Martínková, J. and **Suchánek, M.** (2023) 'Laying foundations for connecting data stewardship domain ontologies', in H. Fujita and G. Guizzi (eds.) *New Trends in Intelligent Software Methodologies, Tools and Techniques*. vol. 371. Amsterdam: IOS, pp. 125–136. Available at: https://doi.org/10.3233/FAIA230229

Martínková, J. and **Suchánek, M.** (2024) 'Towards semantic data management plans for efficient review processing and automation', in *Proceedings of the 13th International Conference on Data Science, Technology and Applications DATA*. vol. 1: Setúbal, Portugal: SciTePress, pp. 543–550. Available at: https://doi.org/10.5220/0012837900003756

Martínková, J., Suchánek, M. and **Pergl, R.** (2024) 'Developing a reference ontouml conceptual model for data management plans: Enhancing consistency and interoperability', in *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. vol. 2. Setúbal, Portugal: SciTePress, pp. 159–166. Available at: https://doi.org/10.5220/0012940000003838

McKain, M.R., Hartsock, R.H., Wohl, M.M. and **Kellogg, E.A.** (2017)' Verdant: Automated annotation, alignment and phylogenetic analysis of whole chloroplast genomes', *Bioinformatics*, 33(1), pp. 130–132. Available at: https://doi.org/10.1093/bioinformatics/btw583

Miksa, T., Walk, P., Neish, P., Oblasser, S., Murray, H., Renner, T., Jacquemot-Perbal, M.C., Cardoso, J., Kvamme, T., Praetzellis, M., Suchánek, M., Hooft, R., Faure, B., Moa, H., Hasan, A. and **Jones, S.** (2021) 'Application profile for machine-actionable data management plans', *Data Science Journal*, 20(32), pp. 1–17. Available at: https://doi.org/10.5334/dsj-2021-032

Milosevic, N. (2016) 'Marvin: Semantic annotation using multiple knowledge sources', arXiv:1602.00515. Available at: https://doi.org/10.48550/arXiv.1602.00515

Müller, H.M., Kenny, E.E. and **Sternberg, P.W.** (2004) 'Textpresso: An ontology-based information retrieval and extraction system for biological literature', *PLOS Biology*, 2(11), p. e309. Available at: https://doi.org/10.1371/journal.pbio.0020309

Nightingale, A. (2009) 'A guide to systematic literature reviews,' *Surgery*, 27(9), pp. 381–384. Available at: https://doi.org/10.1016/j.mpsur.2009.07.005

Nunes, T., Campos, D., Matos, S. and **Oliveira, J.L.** (2013) 'BeCAS: Biomedical concept recognition services and visualization', *Bioinformatics*, 29(15), pp. 1915–1916. Available at: https://doi.org/10.1093/bioinformatics/btt317

Oro, E. and **Ruffolo, M.** (2008) 'XONTO: An ontology-based system for semantic information extraction from PDF documents', in *2008 20th IEEE International Conference on Tools with Artificial Intelligence*. The Institute for Electrical and Electronics Engineers, pp. 118–125. Available at: https://doi.org/10.1109/ICTAI.2008.48

Pölzler, T. (2023). Data *Management Plan: Making Morality Impartial: An Experimental Investigation of the Veil of Ignorance*. Available at https://dmponline.dcc.ac.uk/public_plans.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D. and **Goranov, M.** (2003) 'KIM – semantic annotation platform', in D. Fensel, K. Sycara, and J. Mylopoulos (eds.) *The semantic web – ISWC 2003*. Berlin, Heidelberg: Springer, pp. 834–849. Available at: https://doi.org/10.1007/978-3-540-39718-2_53

RDFa Working Group. (2013) *RDF in attributes (RDFa)*. Available at: https://www.w3.org/2001/sw/wiki/RDFa (Accessed: 17 July 2023).

Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H. and **Jimeno, A.** (2008) 'Text processing through web services: Calling Whatizit', *Bioinformatics*, 24(2), pp. 296–298. Available at: https://doi.org/10.1093/bioinformatics/btm557

Reeve, L.H. and **Han, H.** (2007) 'CONANN: An online biomedical concept annotator', in S. Cohen-Boulakia and V. Tannen (eds.) *Data integration in the life sciences*. Berlin, Heidelberg: Springer, pp. 264–279. Available at: https://doi.org/10.1007/978-3-540-73255-6_21

Rindflesch, T.C., Tanabe, L., Weinstein, J.N. and **Hunter, L.** (1999) 'EDGAR: Extraction of Drugs, Genes and Relations from the biomedical literature', *Biocomputing 2000*, pp. 517–528. Available at: https://doi.org/10.1142/9789814447331_0049

Rocktäschel, T., Weidlich, M. and **Leser, U.** (2012) 'ChemSpot: A hybrid system for chemical named entity recognition', *Bioinformatics*, 28(12), pp. 1633–1640. Available at: https://doi.org/10.1093/bioinformatics/bts183

Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C. and **Chute, C.G.** (2010) 'Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications', *JAMIA*, 17(5), pp. 507–513. Available at: https://doi.org/10.1136/jamia.2009.001560

Settles, B. (2005) 'ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text', *Bioinformatics*, 21(14), pp. 3191–3192. Available at: https://doi.org/10.1093/bioinformatics/bti475

Shakya, A., Wuwongse, V., Takeda, H. and **Ohmukai, I.** (2007) 'OntoBlog: Linking ontology and blogs', *Proceedings of the Semantic Authoring, Annotation and Knowledge Markup Workshop, SAAKM, 2007*. Whistler, BC, Canada, 28–31 October.

Smale, N., Unsworth, K., Denyer, G. and **Barr, D.** (2018) 'The history, advocacy and efficacy of data management plans', *bioRxiv*. Available at: https://doi.org/10.1101/443499

Soavi, M., Zeni, N., Mylopoulos, J. and **Mich, L.** (2020) 'ContracT – from legal contracts to formal specifications: Preliminary results', in J. Grabis and D. Bork (eds.) *The practice of enterprise modeling*. Cham: Springer, pp. 124–137. Available at: https://doi.org/10.1007/978-3-030-63479-7_9

Tanenblatt, M.A., Coden, A. and **Sominsky, I.L.** (2010) 'The ConceptMapper approach to named entity recognition', in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias (eds.) *Proceedings of the seventh international conference on language resources and evaluation, LREC, 2010*. Valletta, Malta, May. European Language Resources Association, pp. 546–51. Available at: https://aclanthology.org/L10-1000/

Tseytlin, E., Mitchell, K., Legowski, E., Corrigan, J., Chavan, G. and **Jacobson, R.S.** (2016) 'NOBLE – Flexible concept recognition for large-scale biomedical natural language processing', *BMC Bioinformatics*, 17(1), p. 32. Available at: https://doi.org/10.1186/s12859-015-0871-y

Vlachidis, A. and **Tudhope, D.** (2016) 'A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain', *Journal of the Association for Information Science and Technology*, 67(5), pp. 1138–1152. Available at: https://doi.org/10.1002/asi.23485

Webster, J. and **Watson, R.T.** (2002) 'Analyzing the past to prepare for the future: Writing a literature review', *MIS Quarterly*, 26(2), pp. xiii–xxiii. Available at: https://www.jstor.org/stable/4132319.

Wohlin, C. (2014) 'Guidelines for snowballing in systematic literature studies and a replication in software engineering', in *Proceedings of the 18th international conference on evaluation and assessment in software engineering, EASE, 2014*. London, England, 13–14 May. New York, NY: Association for Computing Machinery, pp. 1–10. Available at: https://doi.org/10.1145/2601248.2601268

TO CITE THIS ARTICLE:
Martínková, J., Suchánek, M. and Kroha, P. 2025 Automated Semantic Annotation of Data Management Plans: A Systematic Review. *Data Science Journal*, 24: 16, pp. 1–14. DOI: https://doi.org/10.5334/dsj-2025-016

Submitted: 19 January 2025
Accepted: 22 April 2025
Published: 08 May 2025

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