Data Management Plans and Linked Open Data:

exploiting machine actionable data management plans through Open Science Graphs

Elli Papadopoulou[1[0000-0002-0893-8509]] , Alessia Bardi[2[0000-0002-1112-1292]], George Kakaletris[1[0000-0002-2095-1220]], Diamadis Tziotzios[1[0000-0003-1670-4611]], Paolo Manghi[2[0000-0001-7291-3210]] and Natalia Manola[1[0000-0002-3477-3082]]

[1] Athena Research and Innovation Center, 6 Artemidos, 15125 Athens, Greece
[2] Consiglio Nazionale delle Ricerche, 1 Via Moruzzi
56124, Pisa, Italy

Abstract. Open Science Graphs (OSGs) are scientific knowledge graphs representing different entities of the research lifecycle (e.g. projects, people, research outcomes, institutions) and the relationships among them. They present a contextualized view of current research that supports discovery, re-use, reproducibility, monitoring, transparency and omni-comprehensive assessment. A Data Management Plan (DMP) contains information concerning both the research processes and the data generated and/or re-used during a project's lifetime. However, they are missing automated solutions and workflows to connect DMPs with the actual data and other contextual information (e.g., publications, fundings). In an open and FAIR-enabling research ecosystem that is currently being realised, information linking between research processes and research outputs is essential. Argos tool for data management planning directly contributes to the OpenAIRE RG and utilises its underlying services and sources to progressively automate validation and monitoring of Research Data Management (RDM) practices.
This paper provides insight to Argos DMP service integrations with OpenAIRE and the Research Data Alliance (RDA) that populate RGs with DMP entities and relationships and strengthen information exchange in a standard way.

Keywords: Research Data Management; Data Management Plans; Research Graphs; Open Knowledge Graphs.

1 Introduction

Funders' and institutions' mandates on open and FAIR[1][7] data management resulted in new demands for responsible research conduct, affecting traditional practices followed by stakeholders and posing/ claiming a dynamic cultural shift in science, globally. In this effort, research data services are built or reformed to apply and/ or to enable best practices for responsible data management to be performed. In Europe, responsible data management is considered a vital part of Open Science incorporated in Framework

[1] Findable, Accessible, Interoperable, Reusable

Programmes (Horizon2020, HorizonEurope), policy documents (Recommendation on access to and preservation of scientific information)[13] and implementation actions (European Open Science Cloud)[2]. Researchers and students whose grants or graduations heavily rely on compliance with Research Data Management (RDM) mandates are expected to embrace and thrive in the new paradigm of Open Science. In this context, one way RDM can be supported is through the use of tools that facilitate the writing of Data Management Plans (DMPs). DMPs hold information about novel datasets (including raw and auxiliary data) that are produced, but also about data that are reused for the purpose of a scientific mission/ project. DMPs offer valuable documentation regarding how data have been handled, processed, curated, published and preserved throughout a data management lifecycle and therefore serve as the bridge to reproducibility of research and reusability of data by other researchers and interested parties, such as SMEs, easing their further and long-term exploitation.

Open Science Graphs (OSGs) are scientific knowledge graphs representing different entities of the research lifecycle (e.g. projects, people, research outcomes, institutions) and the relationships among them. They present a contextualized view of current research that supports discovery, re-use, reproducibility, monitoring, transparency and omni-comprehensive assessment [1]. They provide insight on the wealth of information in Open Science environments which, according to RDA, can be classified to at least the following subsets: thematic graphs, citation graphs and monitoring graphs[2]. Examples of Open Science Graphs are Open Citations[3], which focuses on bibliographic and citation data; Research Graph[4], which connects scientific literature, research data, project grants and researchers; the Open Research Knowledge Graph, which focuses on research papers and the knowledge available in their full-texts; the FREYA PID Graph[5], which includes interlinked entities provided they are assigned a persistent identifier; and the OpenAIRE Research Graph (OpenAIRE RG)[6][5], which includes descriptive metadata, provenance metadata and links between research results of any kind (literature, software, datasets, and other), organisations, grants, research communities and infrastructures.

Capturing DMPs in OSGs unlocks new potentials, from discovery to curation and contextualisation of information that is carried out in their context. OSGs are enriched with new entities and new relationships that reveal links between DMPs and other entities, such as datasets or metadata, depending on the depth and detail that the respective OSGs goes into. In turn, having those processes in place, provides a good base for monitoring DMPs and assessing their impact as they form a rich resource for exploring RDM trends, e.g. how RDM is perceived and practiced at a given moment in time, about services and how they contribute to RDM best practices and to literacy of researchers etc.

---

[2] Open Science Graphs for FAIR data Interest Group: https://www.rd-alliance.org/groups/open-science-graphs-fair-data-ig

[3] https://opencitations.net/

[4] https://researchgraph.org/

[5] https://www.project-freya.eu/en

[6] https://graph.openaire.eu

2 Argos and the OpenAIRE Research Graph

In the open and FAIR-enabling research ecosystem that is being realised, information linking between research processes and outputs is essential. A DMP contains information concerning the terms and means under which data are utilized or generated throughout a project's lifetime, but lacks automated solutions and workflows that connect it to other useful information such as publications and funding information, thus enabling the creation of complete and coherent research information entities. This is where Argos DMP service[7] [6] thrives by generating machine-actionable DMP outputs (ma-DMPs) [12]. Also, because Argos directly contributes to one of the existing OSGs, the OpenAIRE RG, and utilises OpenAIRE services for bringing together all the bits of information, aiming to progressively automate the processes that contribute to enhancing the information integrity and the validity of European research.

There is rapid development of DMP tools, deployed by funders or institutions, that embed DMPs in RDM lifecycles and offer support and training throughout the process of data management and documentation. Those tools differ at conceptual (e.g. target audience, functionalities, cost) and/ or technical (e.g. data model, technology) levels. Despite aforementioned differences and deviations, most DMP tools are gradually moving towards applying the Research Data Alliance (RDA) standard[8] (RDA standard hereinafter) for interoperability and machine-actionability with successful examples being, among others but not limited to Argos, Data Stewardship Wizard[9] and the DMPonline[10].

The long term machine-actionability vision of Argos encompasses a wide range of other features and is based on standards and interoperability. Among those features are: (semi)automated realisation of DMP terms and statements, such as publication or posting of data into designated repositories, along with corresponding licensing and terms information; validation of DMP claims, such as compliance with metadata standards, presence in repositories, access terms etc; awareness of contributors and authors by notifications and approval and many more.

Focusing on cultivating responsible data professionals and researchers and in support of RDM mandates, OpenAIRE and EUDAT CDI[11] joined efforts to deliver an open platform for Data Management Planning, the OpenDMP software[12], that addresses prevailing requirements and challenges. Argos (argos.openaire.eu) is an instance of the OpenDMP open source software, configured in the OpenAIRE ecosystem, and is available through the OpenAIRE Service catalogue[13] and the European Open Science Cloud (EOSC)[14]. Argos assists the Open Science and FAIR RDM community by applying

7 https://argos.openaire.eu/splash/
8 https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard
9 https://ds-wizard.org/
10 https://dmponline.dcc.ac.uk/
11 https://eudat.eu/eudat-cdi
12 https://gitlab.eudat.eu/dmp/OpenAIRE-EUDAT-DMP-service-pilot/-/tree/master; Wiki page: https://gitlab.eudat.eu/dmp/OpenAIRE-EUDAT-DMP-service-pilot/-/wikis/home
13 http://catalogue.openaire.eu/service/openaire.argos
14 https://providers.eosc-portal.eu/service/openaire.argos

common standards for interoperable, machine-actionable DMPs as defined by the global RDA and by communicating and cooperating with researchers, research communities (e.g. ARIADNEPlus[15]) and funders (e.g. CHIST-ERA[16]) to better reflect on their needs.

Argos features three editors: the DMP Editor and the Dataset editor to accommodate documentation of more than one datasets per DMP, and the Template editor that allows to define personalised structures for DMPs. An example is the H2020 DMPs template [11] that can be used in Argos to create DMPs for projects funded under the European Commission H2020 framework programmes [4].

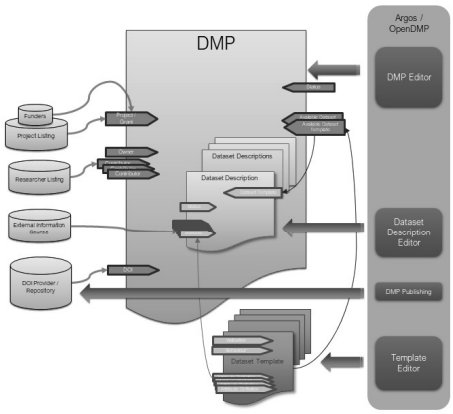

Fig. 1. Argos model in brief

The DMP editor presents the context for the DMPs creation containing questions that address core information about it, such as the DMP scope, the funding behind the given research, responsible actors, etc. The Dataset editor presents the steps and processes followed when managing and handling the data concentrating on questions that show, for example, how Open and FAIR principles have been applied throughout the data lifecycle. That way, Argos becomes more inclusive and flexible as research evolves and new demands arise. It allows datasets to be described separately, per type of data (e.g. a DMP may contain a description of a sensitive dataset and a description of a simulation dataset, etc) and / or per scientific discipline (e.g. a DMP may contain a description of a linguistics dataset and a description of an archaeological dataset, etc), and offers the possibility for descriptions of datasets to be copied and re-used in other DMPs within its platform. The Template editor is visible only to service administrators, who have specific rights to add new templates, modify existing ones and set the rules

[15] The European research community for archaeological research: https://ariadne-infrastructure.eu/

[16] A European network of funding organisations: https://www.chistera.eu/

and conditions for the fields that they include in their templates (e.g. mandatory fields, types of input supported, such as APIs[17], etc). The Template editor is a key enabler of compatibility with the RDA standard as it incorporates a mechanism that maps Argos fields contained in templates to RDA dataset entities. Additionally, Argos allows values in DMPs to be fetched from external APIs like OpenAIRE APIs and ORCID[18] API. Special attention is given to the handling of data that are being re-used by utilising the OpenAIRE API to provide links with the host as well as with the dataset files.

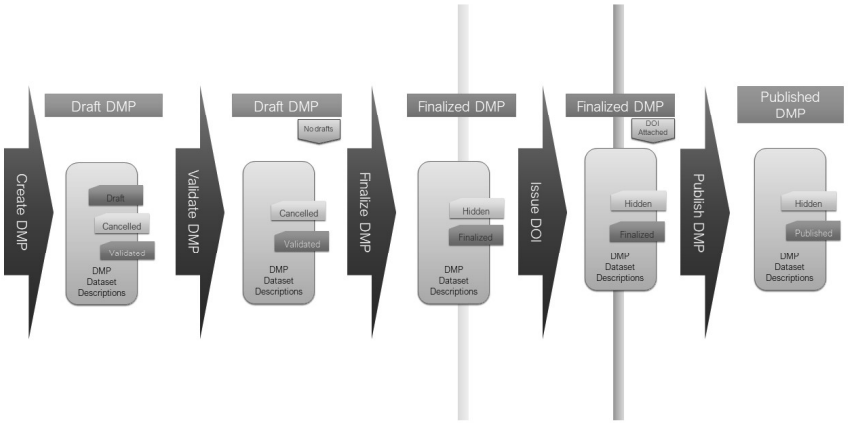

Fig. 2. DMP Lifecycle in Argos

A DMP's lifecycle in Argos consists of a predefined number of internal states followed by versioning and publication mechanisms. A DMP is in draft state while being composed, and while draft it might be either validated (i.e. passing all the requirements of the dataset templates that each of its contained datasets conform to) or non-validated, which is computed substate. Validated DMPs can advance to a finalized state, where editing is restricted, unless the DMP is reverted to draft again. The DMP may be published both internally, i.e. listing of published DMPs offered by the service, or externally. As shown in figure 2, Argos is fully integrated with Zenodo[19], the catch-all repository of OpenAIRE hosted by CERN, and thus offers the option to publish DMPs as outputs in an open and FAIR manner, by assigning DOIs[20] and licenses and by supporting DMPs as living documents through versioning. Once a DOI is issued from Zenodo, reverting a DMP to draft state is not possible and a new version must be created in order to add modifications. Following one cycle of publication, new versions of a

[17] Application Programming Interfaces
[18] ORCID https://orcid.org/
[19] Zenodo https://zenodo.org
[20] Digital Object Identifiers

DMP may be created and linked to their predecessor ones, while the latter continue to survive in the system for reference.

3 Methodology

The process of integrating Argos with the OpenAIRE Research Graph starts with mapping data models between OpenAIRE, Argos and RDA standard and identifying commonalities and differences in the way entities and properties are captured and expressed in each setting. Then, identified gaps and weaknesses are reported and weighted against levels of complexity to address them. Prioritization leads to implementation of new entities and properties in both Argos and the OpenAIRE RG, and to acknowledgment of additional potentials in information exchange between them to be tackled as their joint efforts continue.

Argos applies the RDA standard to deliver DMP outputs that are by design produced in a standard way and can be further exploited in other RDA-compliant DMP platforms. By doing so, it prevents information loss and facilitates a seamless/ uninterrupted research conduct that maximises researchers freedom to switch between DMP providers throughout a research lifecycle. In applying the standard, a mapping activity of entities and properties was performed between Argos data model and the RDA standard. The outcome of this activity was then further examined against the OpenAIRE ecosystem and complemented with mapping to the OpenAIRE data model [5]. The aim was to understand what information is already captured by the RG and how Argos enriches it according to the standard, but also independent of it.

The mapping between the three models has been analysed to identify the level of alignment of the models and understand if any of the models could not accommodate information available in the others. The gap analysis can be considered as the starting point for planning possible updates to the model to ensure interoperability and interlinking of DMPs in the OpenAIRE Research Graph.

The OpenAIRE RG already includes metadata about DMPs that is made available by few content providers (mainly Zenodo) as records compliant to the OpenAIRE guidelines [9,10]. The mappings are also analysed to report about possible new relationships and properties that are not currently included in DMP metadata records but that are instead available in machine actionable DMPs like those produced by Argos.

Of particular interest is the RDA Hackathon[21] that was organized in May 2020, provided the opportunity for active conversation among DMP providers and the RDA community (Active Data Management Plans IG[22]). Argos winning the Hackathon accelerated advancements in terms of interoperability and machine-actionability, also with respect to integrations with the OpenAIRE RG.

---

[21] RDA Hackathon on ma-DMPs: https://github.com/RDA-DMP-Common/hackathon-2020. A full report from Argos participation in the Hackathon will be published by the Data Science Journal (https://datascience.codata.org/).

[22] https://www.rd-alliance.org/groups/active-data-management-plans.html

4 First outcomes

The mapping activity[23] between Argos, OpenAIRE RG and the RDA standard models revealed common entities, properties, and relationships, while observed for major deviations in cardinality to be avoided.

Also, it should be noted that, in order to conform to the needs of ma-DMP fixed schema, without losing the versatility of its templating mechanism, Argos software follows an approach that engages an extensible mechanism for attaching export format converters (ma-DMP being one of them) and semantic tagging of template elements that can be used "at-will" by those converters. The ma-DMP converter makes use of its knowledge of the fixed part of Argos data model as well as attributes attached to various dataset description fields in order to pick the data required for a ma-DMP file.

4.1 Mapping between OpenAIRE RG and RDA standard

The analysis of the mapping between the model of OpenAIRE and the RDA standard shows that:

- Most of the ma-DMP entities can be mapped directly to the OpenAIRE RG (see fig. 3).
- The following properties of the DMP entity can be directly mapped into the OpenAIRE Research Graph: identifier (DMP_id), title, date of creation and modification, description.
- Contact and contributor can be mapped into the OpenAIRE RG with the loss of email information.
- There are entities that are completely absent in the OpenAIRE model. These are: cost indicating total expenditure for data management; metadata schemas showing how datasets are described.
- The dataset entity, including its properties, can not directly fit in the OpenAIRE RG due to data type, cardinality or missing information.
- There is a need for new entities, properties and relationships to be included in the RG thus facilitating information exchange with ma-DMPs.

[23] A close comparison of mapping entities and properties can be found at: https://tinyurl.com/yyakntc9.

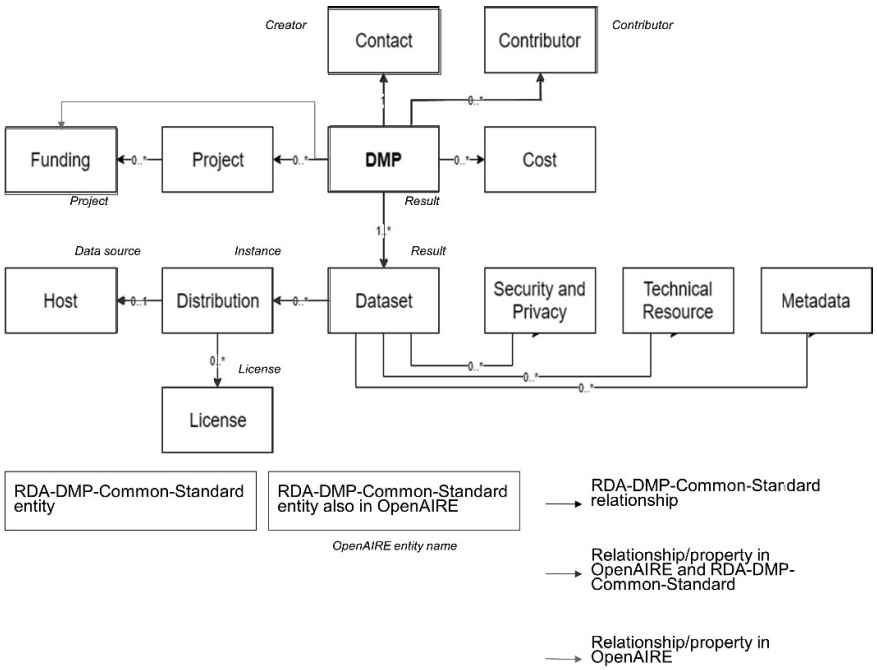

Fig. 3. RDA standard and the OpenAIRE Research Graph Model

The OpenAIRE RG already includes metadata about DMPs that are made available by few content providers (mainly Zenodo) as records compliant to the OpenAIRE guidelines. Metadata includes properties and relationships that are useful for findability (persistent identifiers), accessibility (landing pages and download URLs of different versions of DMPs, access right information), citation and discovery (bibliographic metadata properties), and for tracking (links to funding grants and involved organisations). For the proper typing of DMPs, OpenAIRE exploits the COAR vocabulary which introduced resource type "DMP"[24] in the latest release. Zenodo also introduced a new type of resource so that users can select the type "Data Management Plan" upon deposition. While DMPs were used to be "disguised as project deliverables", the proper typing enables the clear identification of DMPs on Zenodo and on the OpenAIRE Research Graph. Thanks to the adoption of the COAR vocabulary in the COAR network, it is expected that the number of DMPs available in OpenAIRE will grow in the next months.

The main gap can be identified in the different granularity used for the project entity in the OpenAIRE model and the project entity in Argos and the RDA standard. In the OpenAIRE model the project entity represents a grant, which corresponds to the Argos and RDA standard entity called Funding. The Project entity intended as in Argos and RDA standard (i.e. a research activity that can be funded by several grants) is not

[24] http://vocabularies.coar-repositories.org/pubby/resource_type/c_ab20.html

present in the OpenAIRE data model. The OpenAIRE Research Graph features a dedicated relation type between research products and project grants: "isProducedBy/produces" (a product is produced by a project; a project produces research product). This type of relationship can exist between projects and research products of any type (publications, datasets, software and other types of research products), and is currently used by OpenAIRE to model the association between DMPs and project grants. Considering the importance assigned by funders to DMPs (e.g. for H2020 grants, the DMPs is a "living deliverable" that must be updated frequently during the lifetime of the project), and to support the development of added-value services on top of machine actionable DMPs, the OpenAIRE Research Graph will include a relationship with a dedicated semantics "hasDMP"/"hasProject". Although the label "hasProject" is inspired from the RDA ontology [3], the relationship will be re-defined because the domain and range do not correspond, as previously discussed.

As shown also in figure 3, the DMP cost information is not present in the OpenAIRE data model. Cost information is not crucial to improve the level of FAIRness of DMPs or datasets, but it might enable analysis about data management costs useful to funders, organisations, and project administrators.

Interestingly, the DMPs that are currently available in the OpenAIRE Research Graph do not include explicit links to the datasets they refer to. Clearly, such links may not exist when the first version of a DMP is published - because the datasets may not yet exist -, but the expectation is that the datasets are made available during the lifetime of the project. Thanks to the different versions of the DMPs, therefore, it would be possible for OpenAIRE to add relationships with specific semantics between a DMP and the referred datasets. The relationships available in the OpenAIRE Research Graph, inspired from CERIF [8] and Datacite[25], do not include a specific semantics that could depict the association between a DMP and its datasets. OpenAIRE is therefore planning to add a new relationship with semantics "hasDataset", drawn from the core ontology of the RDA standard [3], and a corresponding inverse relationship, which is instead not defined in the standard, to link datasets to the DMPs ("hasDMP").

A final potential for OpenAIRE is identified in the Dataset metadata available in ma-DMPs. Dataset metadata includes information about ethics, security, quality, and preservation that are not currently considered in the OpenAIRE guidelines[9,10] and, therefore, in the OpenAIRE data model. Those kinds of information are potentially useful to studies about responsible research and may be integrated to serve specific analysis or use cases related to RRI (Responsible Research & Innovation) monitoring.

4.2 Mapping between Argos and RDA standard

The mapping highlighted the need for updates in Argos in order to be fully compliant to the RDA standard and strengthen its ma-DMPs. These are:

- Contact, to expose email information of the DMP creator
- Cost, to provide information about expenditure per dataset and DMP (total)

[25] https://support.datacite.org/docs/relationtype_for_citation

- Language, to include in DMPs and datasets
- License, to accommodate access rights imposed by start_date property

Additional observations with respect to Argos are:

- Host property is available via authoritative sources
- Accommodates some entities on the level of datasets and not on the level of DMPs, e.g. ethical_issues_exist

Moreover, a topic that occupied the attention of the Argos was around pre-filling of DMPs, especially focusing on the handling of information concerning re-used datasets in DMP templates. This is a complex and sensitive subject which displays differences according to:

i. the purpose and context of the DMP creation: which most of the time fulfil requirements set in RDM policies enforced by funders or institutions. An example could be institutions integrating their own services on a DMP template and pre-filling it with information inferred by these sources as per the given institution's RDM policy. A common example in RDM policies is the definition of the institutional repository where the datasets will reside. However, even in that case, there are dependencies concerning, among other things, the model of the DMP tool used (for allowing information to be inferred in a template) and the type of data documented (new or re-used datasets).

ii. the time of the DMP lifecycle when the pre-filling takes place: it is inevitable that DMPs at the start of a scientific mission/ project have less information to offer than at the course of the mission/ project when data activities are underway, or even at the end when activities are completed and concrete outputs have been derived. There are differences between new versus re-used data. Existing datasets can be easily claimed by a repository or registry and return back metadata about their title, authors, formats, licenses assigned to them, etc. This information can be pre-filled in the DMP tool and respective institutional DMP templates from inferred sources. On the other hand, this is not the case for new datasets which have not been described or deposited following Open and FAIR practices, yet.

5 Conclusion

The paper aimed to provide a case study of how machine actionable DMPs can be contextualised and exploited in a Linked Open Data environment. The LOD environment used for this activity was the OpenAIRE Research Graph. Argos, the OpenAIRE DMP service, contributed to this work by paving the way to exposing ma-DMPs in OpenAIRE.

The paper explained how Argos is structured, how it applies the RDA standard to increase interoperability and machine-actionability of its outputs, how it integrates with other services to increase openness and FAIRness of its output, to finally show how it operates in the OpenAIRE ecosystem and how it enriches the OpenAIRE RG. This is an ongoing effort between Argos and the OpenAIRE RG, supported thanks to the OpenAIRE API providing information about organizations, data sources and datasets,

obtained by collecting metadata records from more than 12K trusted scholarly communication sources (including Datacite, Crossref, re3data, OpenDOAR, Grid.ac).

Work highlighted the areas needed to be strengthened in both Argos and OpenAIRE in order to accommodate the RDA standard as well as to populate the RG with DMP entities and create links with other outputs. Observations made during the mapping activity of entities and properties between the three data models of Argos, OpenAIRE and RDA showed that when direct mapping couldn't be fulfilled, information might still be able to be found in more abstract/ general fields of OpenAIRE or Argos though diverged in cardinality and / or data type; some information might still be covered by Argos DMP outputs as they enter the OpenAIRE RG or tweaked to accommodate the needs of maDMPs documentation or, rarely, they may be omitted in information exchange. OpenAIRE, also, highlighted the value added in contextualising ma-DMPs as they contain structured and specialised information, especially about datasets, which can not be found in original / traditional DMP documents. Searching OpenAIRE for DMP outputs outside Argos, showed that DMPs are still typed as generic publications or reports, and not as data management plans. The introduction of a specific term for DMPs in global vocabularies, such as COAR's, is expected to improve the current status, although the adoption of the new term may take some time to be wide-spread.

References

1. A. Aryani, M. Fenner, P. Manghi, A. Mannocci and M. Stocker, "Open Science Graphs Must Interoperate!", 24th International Conference on Theory and Practice of Digital Libraries (TPDL), Lyon, France, 2020. doi: 10.1007/978-3-030-55814-7_16
2. European Commission. (2016). European Cloud Initiative - Building a competitive data and knowledge economy in Europe. Luxembourg: Office for Official Publications of the European Communities.
3. Fajar J. Ekaputra,João Cardoso,Leyla Garcia,Marie Christine Jacquemot. The standard Ontology. Retrieved from: https://w3id.org/dcso/ns/core/3.0.2
4. Data Management, European Commission Webpage, https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm, last accessed 2020/08/17
5. Manghi, Paolo, Bardi, Alessia, Atzori, Claudio, Baglioni, Miriam, Manola, Natalia, Schirrwagen, Jochen, & Principe, Pedro. (2019, April 17). The OpenAIRE Research Graph Data Model (Version 1.3). Zenodo. http://doi.org/10.5281/zenodo.2643199
6. Papadopoulou, Elli, Kakaletris, Georgios, Tziotzios, Diamadis, Moa, Hanne, & Hasan, Adil. (2020). ARGOS: a collaborative tool to plan and follow your data. Zenodo. http://doi.org/10.5281/zenodo.3898249
7. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
8. CERIF data model: https://eurocris.org/services/main-features-cerif
9. Houssos, Nikos, et al. "OpenAIRE guidelines for CRIS managers: supporting interoperability of open research information through established standards." Procedia Computer Science 33 (2014): 33-38.
10. Schirrwagen, Jochen, & Baglioni, Miriam. (2018). OpenAIRE Guidelines for institutional and thematic repository managers 4.0 (Version 4.0.0). Zenodo.

11. European Commission. Template Horizon 2020 Data Management Plan (DMP). https://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotated_en.pdf
12. Miksa T, Simms S, Mietchen D, Jones S (2019) Ten principles for machine-actionable data management plans. PLOS Computational Biology 15(3): e1006750. https://doi.org/10.1371/journal.pcbi.1006750
13. Official Journal of the European Union. COMMISSION RECOMMENDATION (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information. [online] Available at: https://eur-lex.europa.eu/legalcontent/EN/TXT/PDF/?uri=CELEX:32018H0790&from=EN