Architecture and Planning Journal (APJ)

Volume 28 Issue 3 *ASCAAD 2022 - Architecture in the Age of the Metaverse – Opportunities and Potentials*
ISSN: 2789-8547

Article 40

March 2023

RESEARCH DATA MANAGEMENT AND A SYSTEM DESIGN TO SEMI-AUTOMATICALLY COMPLETE INTEGRATED DATA MANAGEMENT PLANS [POSITION PAPER]

SYED ASHFAQ HUSSAIN SHAH
Chair of Architectural Informatics, Technical University of Munich, Germany, syed.hussain@tum.de

FRANK PETZOLD
Chair of Architectural Informatics, Technical University of Munich, Germany, petzold@tum.de

Recommended Citation

HUSSAIN SHAH, SYED ASHFAQ and PETZOLD, FRANK (2023) "RESEARCH DATA MANAGEMENT AND A SYSTEM DESIGN TO SEMI-AUTOMATICALLY COMPLETE INTEGRATED DATA MANAGEMENT PLANS [POSITION PAPER]," *Architecture and Planning Journal (APJ)*: Vol. 28: Iss. 3, Article 40.
DOI: https://doi.org/10.54729/2789-8547.1235

RESEARCH DATA MANAGEMENT AND A SYSTEM DESIGN TO SEMI-AUTOMATICALLY COMPLETE INTEGRATED DATA MANAGEMENT PLANS [POSITION PAPER]

Abstract

Data is an integral part of modern scientific work. Good research data management (RDM) and the communication of the related information is extremely an important matter. It is not only crucial for the ongoing research and its claims but also for the future uses of data. In recent years some guiding principles, e.g. FAIR principles and initiatives at the national and international level, e.g. NFDI, NFDI4Ing have also been founded to improve RDM. The data and its metadata are often handled in file system like structures which are versioned and logged. The information relating to the data handling are documented in data management plan (DMP). DMPs are also usually managed in similar file structures. These are made available in editable document formats as well as online free-text editable forms to which users are required to keep updating manually. These are isolated documents which have neither direct relation to data for verification nor are common to understand with consistency. In this paper, research data management of large-scale interdisciplinary projects is presented. On one hand it introduces, contemporary practices of RDM and on the other hand it helps researchers to determine the features of RDM system in the situations when it comes to select or develop a system for the same purpose. It further introduces a system design for semi-automatic completion of DMP functions in collaborative environment a.k.a. virtual research environment (VRE). It is assumed that the proposed system will assist and enable users to update semi-automatically integrated DMP during all phases of data life cycle. Direct relation to the data for verification, common understanding and consistency will also be maintainable.

Keywords

Research data management (RDM), dynamic Data management plan (dDMP), Virtual research environment (VRE), Research data management system, Open science.

RESEARCH DATA MANAGEMENT AND A SYSTEM DESIGN TO SEMI-AUTOMATICALLY COMPLETE INTEGRATED DATA MANAGEMENT PLANS [POSITION PAPER]

SYED ASHFAQ HUSSAIN SHAH AND FRANK PETZOLD

Chair of Architectural Informatics, Technical University of Munich, Germany
syed.hussain@tum.de
petzold@tum.de

ABSTRACT

Data is an integral part of modern scientific work. Good research data management (RDM) and the communication of the related information is extremely an important matter. It is not only crucial for the ongoing research and its claims but also for the future uses of data. In recent years some guiding principles, e.g. FAIR principles and initiatives at the national and international level, e.g. NFDI, NFDI4Ing have also been founded to improve RDM. The data and its metadata are often handled in file system like structures which are versioned and logged. The information relating to the data handling are documented in data management plan (DMP). DMPs are also usually managed in similar file structures. These are made available in editable document formats as well as online free-text editable forms to which users are required to keep updating manually. These are isolated documents which have neither direct relation to data for verification nor are common to understand with consistency. In this paper, research data management of large-scale interdisciplinary projects is presented. On one hand it introduces, contemporary practices of RDM and on the other hand it helps researchers to determine the features of RDM system in the situations when it comes to select or develop a system for the same purpose. It further introduces a system design for semi-automatic completion of DMP functions in collaborative environment a.k.a. virtual research environment (VRE). It is assumed that the proposed system will assist and enable users to update semi-automatically integrated DMP during all phases of data life cycle. Direct relation to the data for verification, common understanding and consistency will also be maintainable.

Keywords: Research data management (RDM), dynamic Data management plan (dDMP), Virtual research environment (VRE), Research data management system, Open science.

ملخص

تعتبر البيانات جزءا لا يتجزأ من العمل العلمي الحديث. كما تعتبر الإدارة الجيدة لبيانات البحث وإيصال المعلومات ذات الصلة أمرًا بالغ الأهمية. وليس ذلك مهمًا فقط للبحث المستمر ولكن أيضًا للاستخدامات المستقبلية للبيانات. وقد ظهرت فى السنوات الأخيرة بعض المبادئ التوجيهية، على سبيل المثال مبادئ ومبادرات FAIR على المستويين الوطني والدولي، كما تم تأسيس NFDI و NFDI4Ing أيضًا لتحسين إدارة بيانات البحث. غالبًا ما يتم التعامل مع البيانات والبيانات الوصفية الخاصة بها في نظام الملفات مثل هياكل بيانات يتم إصدارها وتسجيلها. ويتم توثيق المعلومات المتعلقة بمعالجة البيانات في خطة إدارة البيانات (DMP). تتم أيضًا إدارة DMPs عادةً في هياكل ملفات مماثلة، ويتم توفيرها في تنسيقات المستندات القابلة للتحرير بالإضافة إلى نماذج النصوص المجانية القابلة للتحرير عبر الإنترنت والتي يُطلب من المستخدمين تحديثها يدويًا. هذه مستندات معزولة ليس لها علاقة مباشرة بالبيانات من أجل التحقق ولا من الشائع فهمها باتساق. في هذه الورقة البحثية، يتم تقديم إدارة بيانات البحث للمشاريع متعددة التخصصات واسعة النطاق. من ناحية، يقدم هذا الطرح ممارسات معاصرة لـ RDM ومن ناحية أخرى يساعد الباحثين على تحديد ميزات نظام RDM في المواقف عندما يتعلق الأمر باختيار أو تطوير نظام للغرض نفسه. يقدم أيضًا تصميم نظام للإكمال شبه التلقائي لوظائف DMP في بيئة تعاونية مثل بيئة البحث الافتراضية (VRE). من المفترض أن النظام المقترح سيساعد المستخدمين ويمكّنهم من تحديث برنامج إدارة البيانات المتكاملة شبه التلقائي خلال جميع مراحل دورة حياة البيانات. كما يمكن الحفاظ على العلاقة المباشرة ببيانات التحقق والفهم المشترك والاتساق.

الكلمات المفتاحية: إدارة بيانات البحث، خطة ديناميكية لإدارة البيانات، بيئة البحث الافتراضية، نظم إدارة بيانات البحث، العلم المفتوح.

1. INTRODUCTION

Scientific investigations are not only about theory and experiments but also include procedures and conducts (The Royal Society, 2012, Easterbrook, 2014, Leek and Peng, 2015). Communication of practices and results contributes together to reproduce and replicate the very same exercise again and again. Importance of reproducible findings have long been rooted in the modern investigation methodologies, i.e. science (The Royal Society, 2012). Whereas, digital instrumentation at large scale and dependence on data and computing machinery have reinforced the demand for the same in recent years (Baker, 2016, Stodden, 2010). As a result, notions emerge to align quantitative & qualitative matrices and the activities to define frameworks for investigation and scholarly communication (Hicks et al., 2015, DORA, 2012). These frameworks levy actual research activity with intensive disciplines, e.g. software engineering and also require knowledge and compliance with technical details of infrastructures as well as rules and regulations laid down by the concerned authorities, i.e. publication and funding agencies etc. (Wilson et al., 2014, Tenopir et al., 2011, Data Citation Synthesis Group, 2014, Stodden et al., 2013).

FAIR data principles and Open science are the two widely supported movements. FAIR principles set the goals for Findability, Accessibility, Interoperability, and Reuse of data. Whereas Open science not only set goals to make scientific research, data, and dissemination accessible to all levels of an inquiring society but also suggest principles of openness to the whole research cycle (Wilkinson et al., 2016, FOSTER). Thus, a classical research activity gets coupled with intensive tasks of information logistics (Klein, 1993).

However, it is assumed that RDM support services and systems could rescue from complications and relieve workloads of an investigation and its scholarly communication. These systems could achieve this by encapsulating the rules and workflows while offering collaboration, fostering adoptability and interoperability between diverse and multiple sources (Van Gorp and Mazanek, 2011, Candela et al., 2013, Gray et al., 2005).

RDM refers to the activities relating to the storage, organisation, documentation, and dissemination of research data. These are continuous activities which are needed to be performed during the period of the research project (Borghi et al., 2018).

Different systems and services are being developed and proposed to accomplish challenging goals of RDM. In this context data types are defined based on its need and state, i.e., active and non-active research data. Data is considered active when it is part of an active phase of research, where data is considered non-active when it is not part of active phase of research. In this paper approaches and systems dealing with active research data are described from simple to complex in an incremental way.

In following sections, an overview of RDM policy and strategy, research data, a review of approaches together with types of RDM systems to manage research data and a system design for integrated dynamic DMPs are presented.

2. RDM POLICY AND STRATEGY

To ensure best practices during research, an RDM activity starts with the development of RDM policy and strategy as a binding document for all the participants. This document defines research data, its related information together with research outcomes, file formats to communicate research outcomes, naming and versioning conventions, initial metadata standards, platform and information infrastructure, categories of roles and responsibilities of the participants (Jones et al., 2013).

3. RESEARCH DATA

Different bodies define research data differently. However, for the sake of reference in this paper research data described by German Research Foundation (DFG) guidelines is considered as follows.

“Research data includes measurement data, laboratory values, audiovisual information, texts, survey or observation data, methodological test procedures and questionnaires.

Compilations, software, and simulations can equally represent a central result of scientific research and are therefore also included under the term research data. Research data in some subject areas is based on the analysis of objects (such as tissue, material, rock, water and soil samples, test specimens, installations, artefacts, and art objects), so its handling must be just as careful, and consideration must be given to a technically adequate option for subsequent reuse whenever meaningful and possible. Should subsequently reuse of the resulting research data be closely associated with objects, then please also elaborate on this by providing all relevant information. Please consider the existing standards in your discipline, any current subject-specific recommendations and any existing infrastructure services." (DFG, 2021).

4. APPROACHES TO MANAGE RESEARCH DATA

This section first provides an overview of versioning concept. Then introduces versioning techniques and systems for RDM including strategies relating to the storage of data and then about organisation of the data. Lastly, documentation of the data.

4.1. Data Versioning

During research, data evolves and goes through different stages and processes, e.g. from raw data to processed data and to data as product. So, it could have different versions. The stages of data are modelled as data lifecycle (Weber and Kranzlmüller, 2019).

For a modern research it is not only important to present the results but also provide a clear protocol to allow successful repetition and extension (Mesirov, 2010). Thus, maintaining different versions become important for repeatability and reproducibility of the same results. Each new version of data is maintained in a way that it has some important associated information, e.g. date and time of its creation, author, new contents, reference to the older version and description of changes.

4.2. RDM by Means of Versioning

There have been different strategies to maintain different versions of the data. A very basic and classic approach is that on every change a newer file of the data is created and saved with corresponding information. Creating versions in this way could cause creation of lots of files which could become difficult to manage. To deal with this issue software tools and solutions are being developed under the theme of Version Control System (VCS). These systems maintain changes and log associated information of a file or set of files in their internal databases. In this way they offer a simpler organisation of different versions as well as enable users to restore specific versions later (git-scm).

Those software tools are being categorised as Local VCS, Centralised VCS and Distributed VCS based on their design. The key difference among them is as follows. Local VCS maintains all the versions of data on local system. GNU RCS[1] is an example of such a system. Centralised VCS maintains all the versions of data on a central server and users can access specific version using client software. CVS[2] and Apache Subversion[3] are examples of such systems. Distributed VCS maintains data on a server with a difference that every client also has all the history of data versions. Therefore, they do not depend on the server as much as they do in case of Centralised VCS. Git[4] and Mercurial[5] are examples of such systems.

Simple version control systems have limitations. Therefore, to collaborate and reuse data, information managed in simple VCSs are not sufficient.

4.3. RDM Using Systems Built Around Versioning Systems

In case of collaborative research projects in which scientists are from different disciplines, stationed at distinct locations and generating heterogeneous data, the task of RDM requires a comprehensive solution. Because requirements for the management of data

[1] https://www.gnu.org/software/rcs
[2] http://cvs.nongnu.org
[3] https://subversion.apache.org
[4] https://git-scm.com
[5] https://www.mercurial-scm.org

in such cases get diversified, e.g. collaborative work requires comprehensive access management not only for already saved file like data but also for real-time collaborative authoring and viewing, internal communication, commenting and annotation, custom metadata etc. Therefore, applications and services around VCS have also been developed to constitute new range of systems. These systems introduce further features and tools, e.g. unique IDs, wiki pages, project model, project pages, tasks and task boards, metadata creation and extraction, support for standard protocols and APIs (Amorim et al., 2017).

These systems can broadly be categorised in four main categories based on their function during RDM activity: -

- **Storage systems:** These are the range of systems which support to save data and manage access to it. Some of these systems also offer document authoring support. Example of these systems are Powerfolder[6], OwnCloud[7] etc.
- **Development support systems:** These are the range of systems which are developed with the aim to manage code and code alike data, which is meant to exhibit special structure, compile and execute. These systems further extend functionality and offer convenient features, e.g., continuous integration and deployment. GitHub[8], GitLab[9], Overleaf[10] and similar authoring and coding systems are examples of such systems. In some cases, these systems, e.g., GitHub might also be used as publication platforms and repositories.
- **Publication platforms and repositories:** These are the range of systems where data is usually parked at the end of its active state. Data in these systems is usually registered to release and make public. There are exceptions that some of these systems might not support multiple versions of the same data. Thus, each submission might be managed independently. Examples of these systems are mediaTUM[11], Zenodo[12]/ InvenioRDM[13], CKAN[14], DSpace[15], Fedora[16], Dataverse[17] etc. The category of publishing systems may also include indexing platforms, e.g., DataCite[18] where researchers could publish their metadata.
- **Virtual Research Environments (VRE):** These are also named as Science Gateways (SG) and Virtual Laboratories (VL). These are relatively new genre of systems to offer range of tools for complete research workflows. Apart from their own storage system, these systems offer integration and interoperability with external systems and infrastructures to aggregate and disseminate data while offering central point for controls. Example of these systems are eWorkbench[19], VRE4EIC[20], OSF[21] etc. In industry similar genre of systems are taking place under the flagship of Hybrid cloud (Marcio et al., 2017).

[6] https://www.powerfolder.com
[7] https://owncloud.com
[8] https://github.com
[9] https://gitlab.com
[10] https://www.overleaf.com
[11] https://mediatum.github.io
[12] https://zenodo.org
[13] https://inveniosoftware.org
[14] https://ckan.org
[15] https://dspace.lyrasis.org
[16] https://duraspace.org/fedora
[17] https://dataverse.org
[18] https://datacite.org
[19] https://eworkbench.github.io
[20] https://vre4eic.ercim.eu
[21] https://osf.io

4.4. Data Organisation

There are three distinctive ways to organise data in RDM systems, i.e. file based, form based and project based.

- **File based:** In file-based approach data is stored usually under file management system using file objects. In this scheme files are organised under folder or directory hierarchies. Directories are created based upon the themes, e.g., instrument, laboratory, month, topic to organise similar files. For RDM activity a standard file system layout and naming conventions for files are defined which researchers are required to follow. In some cases, it not only includes standard data files but also includes documentation and metadata files. Directories are applied VCS, e.g., Git. In order to maintain folders independent of each other modular approach of VCS, e.g. git submodule are applied (Spreckelsen et al., 2020). Applications like file manager are used to organise and access such data.
- **Form based:** In form-based approach data is accessed, visualised and versioned by the means of custom software programs, e.g. data relating to the events is managed by an appointment software module. Modern systems adopt this approach to maintain and organise documentation and data management plan as well as metadata. In this approach data is stored usually in data management systems, e.g., PostgreSQL[22].
- **Project based:** This approach uses both file and form-based approaches. It is best suited for large projects consisting of multiple hierarchies of sub projects. It uses project model and employs namespaces concepts to maintain contextual aspects of data and information as well as access management. In this approach, project models are defined and programmed to encapsulate information specific organisation/ tools, e.g., appointment and calendar, files and storage, sub projects, metadata, and data management plan.

4.5. Data Documentation

Just like data organisation approaches there have been two basic approaches to document information relating to data and research practices, i.e., file based, and form based. Form based approach has advantages of automation due to being supported by custom software program. Whereas, file-based approach requires researchers to maintain each and every piece of information in files manually. For example, if metadata is maintained in a file, then whenever data is updated, researchers are required to update corresponding fields of metadata in the file manually. However, in case of form-based approach log of the files are automatically maintained through the data file updating process. Advance case of documentation of practices in data management plan is discussed in the next section.

5. SYSTEM DESIGN FOR DYNAMIC DMP

Data Management Plans (DMPs) are described as a key element of good data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a project (European Commission). It is a requirement for all the DFG projects and doctoral theses. It is also one of the key topics of national and international research data initiatives, e.g., NFDI4Ing[23], RDA[24].

DMP consists of the text that describes intentions at the beginning, practices during the investigation process and later the applied approaches that how the data has been handled from first conception of the project till it is archived or, if deemed necessary, deleted.

In the beginning, it is composed by formulating questions. These questions are raised by the funding agencies like DFG, EU or by the competent authorities of the projects or consortiums, e.g., TRR277 AMC[25] for domain specific requirements and then it is offered to

[22] https://www.postgresql.org
[23] https://nfdi4ing.de
[24] https://www.rd-alliance.org/groups/dmp-common-standards-wg
[25] https://amc-trr277.de

their participants. Participants then answers in a free-text form. It is recognised that this approach put an administrative burden on researchers. And the provided information is not as much useful as it should be while considering its anticipated benefits (Smale et al., 2020).

In the following sections, first a review of related work, core features of TUM Workbench as VRE to manage research data, background and rationale are explained, then a system design for semi-automatic completion of DMP functions is introduced.

5.1. Related Work

The review of existing solutions showed that, at the moment it is a simple document which is available in editable document formats as well as online free-text editable forms (EC H2020, 2018, OpenAIRE, 2017, RDMO). In order to maintain common understanding and consistency of the answers, this document is also offered with some examples of answers and hints for its authors (EC H2020, 2018). There are solutions that help authors to create an accustomed DMP document (CLARIN-D, OpenAIRE and EUDAT, RDMO). In addition to create a custom DMP, it goes to make way to link DMP to data archiving system which takes place at the end of the data life cycle (RDMO).

Due to realisation that DMP is an active document, new trends are emerging to make this document machine actionable by defining application profiles and bridging the gap between data and its corresponding DMP. And to facilitate authors to author this document (Miksa et al., 2019, DCC, 2021, Miksa et al., 2017). There have been efforts to define data model for DMP (Miksa et al., 2021b, Freudenberg et al., 2016) then the definition of automated workflows based on business processes in an institutional context (Miksa et al., 2021a).

The work being presented in this paper relates to a collaborative environment, i.e. VRE in which projects and research may belong to different institutions regardless of a specific location. The design and approach presented in this paper has a potential to semi-automatically complete DMP tasks which have not been demonstrated in other approaches so far, e.g., data collection and selection details. The concept of dynamic DMP already pertains to idea of data model. It is defined as a plan which can adopt or suggest user based on the information available on the associated data part and vice versa.

5.2. TUM Workbench

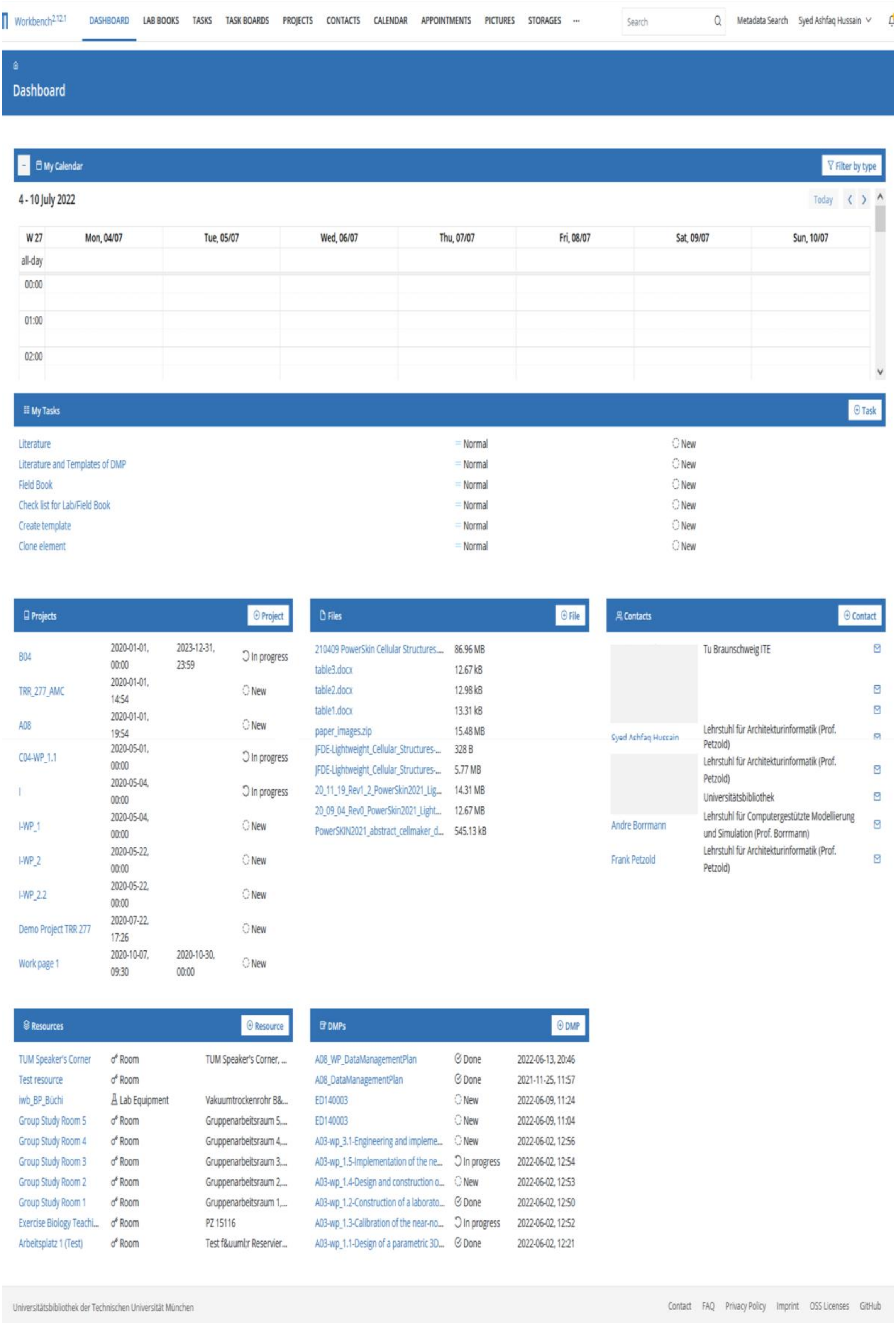
TUM Workbench[26] is an installation of eWorkbench[27] which is being developed by Technical University of Munich Library. Its technological backend including data backup and archive services are being supported by Leibniz-Supercomputing Centre (LRZ).

TUM Workbench offers namespaces concept for user, project, sub project as well as for available tools to manage research data. Thus, hierarchical, and composite data management is possible while maintaining inheritance, contexts and access rights. Tools available in TUM Workbench are included storage, file, appointment, calendar, note/ lab book, task, task board, commenting and annotating, notification as well as communication tools. Tools are termed as elements. Access and rights management are realised by assigning role in a project. Rights can be customised for elements on individual basis too. Information and data can be registered in terms of online forms and files. Apart from basic operations over data, i.e., create, read, update, move, TUM Workbench also offers link/ unlink operation to refer to and to present a simple association between elements, e.g. to associate data and corresponding DMP. Other common features of all the elements include unique ID, versioning, change logs, support for DataCite as basic metadata, addition of custom metadata fields, web GUI, support for WebDAV and CalDAV protocols, REST API, plug-in support, and management of custom DMP templates. Figure 1 presents a snapshot of web GUI based user dashboard in TUM Workbench.

Based on the above-mentioned features figure 2 presents data management model of AMC in TUM Workbench.

[26] https://workbench.ub.tum.de

[27] https://eworkbench.github.io



Literature	Normal	New
Literature and Templates of DMP	Normal	New
Field Book	Normal	New
Check list for Lab/Field Book	Normal	New
Create template	Normal	New
Clone element	Normal	New

B04	2020-01-01, 00:00	2023-12-31, 23:59	In progress
TRR_277_AMC	2020-01-01, 14:54		New
A08	2020-01-01, 19:54		New
C04-WP_1.1	2020-05-01, 00:00		In progress
I	2020-05-04, 00:00		In progress
I-WP_1	2020-05-04, 00:00		New
I-WP_2	2020-05-22, 00:00		New
I-WP_2.2	2020-05-22, 00:00		New
Demo Project TRR 277	2020-07-22, 17:26		New
Work page 1	2020-10-07, 09:30	2020-10-30, 00:00	New

210409 PowerSkin Cellular Structures....	86.96 MB
table3.docx	12.67 kB
table2.docx	12.98 kB
table1.docx	13.31 kB
paper_images.zip	15.48 MB
JFDE-Lightweight_Cellular_Structures-...	328 B
JFDE-Lightweight_Cellular_Structures-...	5.77 MB
20_11_19_Rev1_2_PowerSkin2021_Lig...	14.31 MB
20_09_04_Rev0_PowerSkin2021_Light...	12.67 MB
PowerSKIN2021_abstract_cellmaker_d...	545.13 kB

	Tu Braunschweig ITE
Syed Ashfaq Hussain	Lehrstuhl für Architekturinformatik (Prof. Petzold)
	Lehrstuhl für Architekturinformatik (Prof. Petzold)
	Universitätsbibliothek
Andre Borrmann	Lehrstuhl für Computergestützte Modellierung und Simulation (Prof. Borrmann)
Frank Petzold	Lehrstuhl für Architekturinformatik (Prof. Petzold)

TUM Speaker's Corner	Room	TUM Speaker's Corner, ...
Test resource	Room	
iwb_BP_Büchi	Lab Equipment	Vakuumtrockenrohr B&...
Group Study Room 5	Room	Gruppenarbeitsraum 5,...
Group Study Room 4	Room	Gruppenarbeitsraum 4,...
Group Study Room 3	Room	Gruppenarbeitsraum 3,...
Group Study Room 2	Room	Gruppenarbeitsraum 2,...
Group Study Room 1	Room	Gruppenarbeitsraum 1,...
Exercise Biology Teachi...	Room	PZ 15116
Arbeitsplatz 1 (Test)	Room	Test f&uuml;r Reservier...

A08_WP_DataManagementPlan	Done	2022-06-13, 20:46
A08_DataManagementPlan	Done	2021-11-25, 11:57
ED140003	New	2022-06-09, 11:24
ED140003	New	2022-06-09, 11:04
A03-wp_3.1-Engineering and impleme...	New	2022-06-02, 12:56
A03-wp_1.5-Implementation of the ne...	In progress	2022-06-02, 12:54
A03-wp_1.4-Design and construction o...	New	2022-06-02, 12:53
A03-wp_1.2-Construction of a laborato...	Done	2022-06-02, 12:50
A03-wp_1.3-Calibration of the near-no...	In progress	2022-06-02, 12:52
A03-wp_1.1-Design of a parametric 3D...	Done	2022-06-02, 12:21

Fig.1: Snapshot of a user dashboard in TUM Workbench.

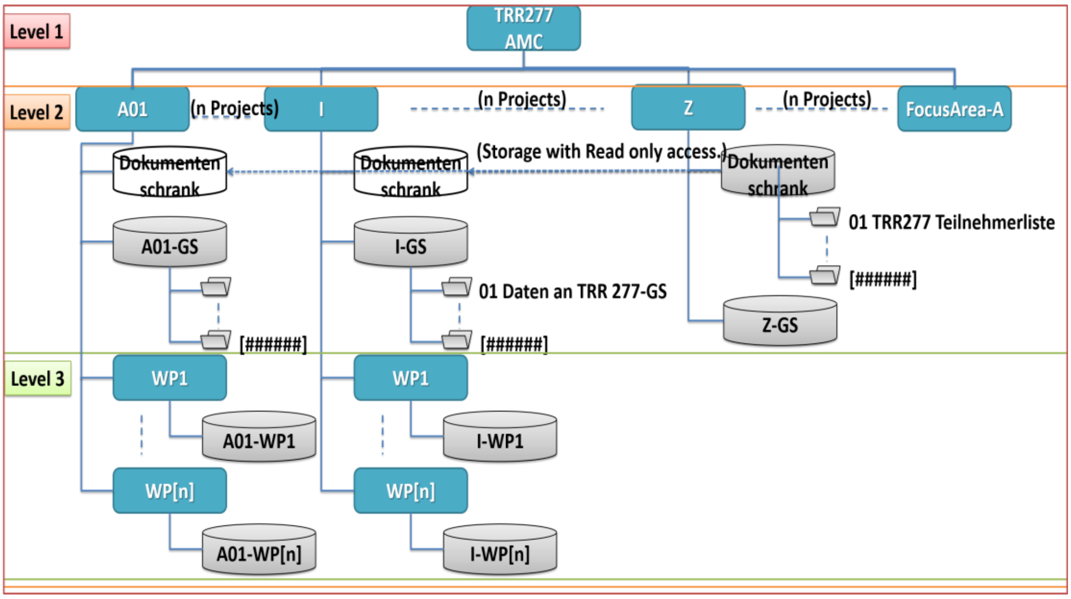


Fig.2: Data management model of AMC in TUM Workbench.

AMC is a trans-regional centre hosted by TU Munich and TU Braunschweig. Whereas the scientists and labs are situated at different locations. The research work of the AMC is grouped in three focus areas, i.e., Area A: Materials and Processes, Area B: Computational Modelling and Control, and Area C: Design and Construction[28].

5.3. Background and Rationale

It is assumed that during research work while interacting with tools, users generate useful information which may be recorded by means of automation or users' manual insertions. For example, when user create a dataset by using upload feature of the VRE, VRE can record date of creation, size, author's name, and contextual information under which data is created etc. And in case of updating existing data users are required to provide change message which can lead to the information about data processing, changes in data etc. All this information collectively constitutes metadata of the corresponding data as shown in figure 3. This very information can be consumed by associated dynamic DMP either to automate its fulfilment without requiring users to type in the form or by suggesting information based on such data to users to semi-automatically fill the form.

As such, it is assumed that metadata models supported research data management plans could offer an opportunity to reduce the time consuming, repetitive manual processing and often error prone work of documenting DMPs. And increase DMPs' acceptance.

5.4. System Design, Components, and Interaction Model

Figure 4 presents system design and respective components of the suggested system. Each component may consist of multiple subcomponents. Following is the explanation of each main component and information flow.

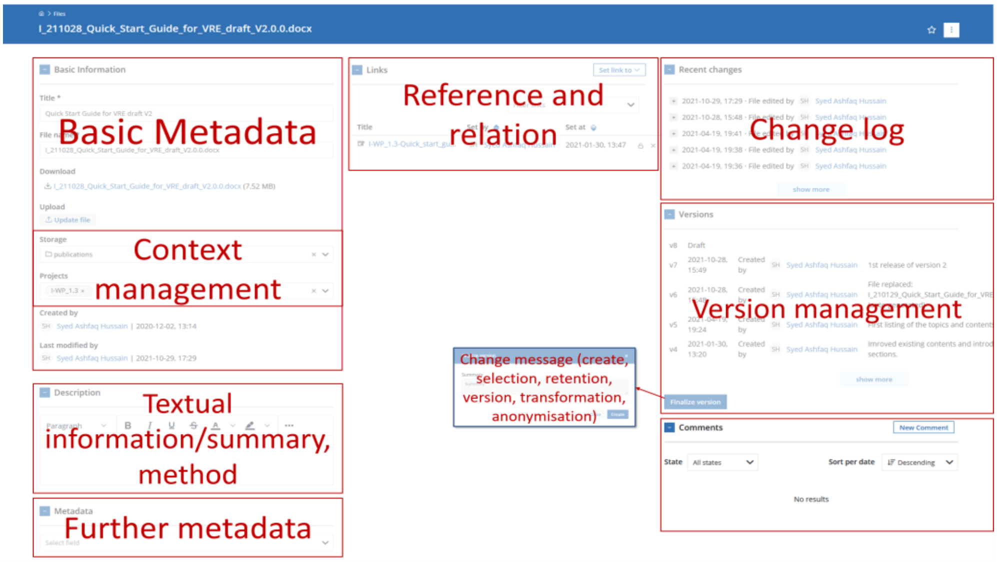
5.4.1. Data

Data component represents dataset with which DMP is associated. In this diagram the hierarchy of the data is presented where all the contextual information may also be retrieved. Depending on the organisation a dataset may consist of single file, storage, project etc.

[28] https://amc-trr277.de/amctrr277-research

5.4.2. Metadata

Every dataset in TUM Workbench has associated metadata. Metadata component represents that metadata part. It may vary depending on the dataset. Even it may vary for the similar types of datasets. It consists of two parts, i.e., form and metadata model.



Files

I_211028_Quick_Start_Guide_for_VRE_draft_V2.0.0.docx

Basic Information

Title *

Quick Start Guide for VRE draft V2

File name

I_211028_Quick_Start_Guide_for_VRE_draft_V2.0.0.docx

Download

I_211028_Quick_Start_Guide_for_VRE_draft_V2.0.0.docx (7.52 MB)

Upload

Update file

Storage

publications

Projects

I-WP_1.3

Created by

SH Syed Ashfaq Hussain | 2020-12-02, 13:14

Last modified by

SH Syed Ashfaq Hussain | 2021-10-29, 17:29

Description

Paragraph B I U S A ...

Metadata

Select field

Links

Set link to

Title | Set by | Set at

I-WP_1.3-Quick_start_gu... | SH Syed Ashfaq Hussain | 2021-01-30, 13:47

Recent changes

- 2021-10-29, 17:29 · File edited by SH Syed Ashfaq Hussain
- 2021-10-28, 15:48 · File edited by SH Syed Ashfaq Hussain
- 2021-04-19, 19:41 · File edited by SH Syed Ashfaq Hussain
- 2021-04-19, 19:38 · File edited by SH Syed Ashfaq Hussain
- 2021-04-19, 19:36 · File edited by SH Syed Ashfaq Hussain

show more

Versions

v8 Draft
v7 2021-10-28, 15:49 Created by SH Syed Ashfaq Hussain 1st release of version 2
v6 2021-10-28, 15:48 Created by SH Syed Ashfaq Hussain File replaced: I_210129_Quick_Start_Guide_for_VRE
v5 2021-04-19, 19:24 Created by SH Syed Ashfaq Hussain First listing of the topics and content
v4 2021-01-30, 13:20 Created by SH Syed Ashfaq Hussain Imroved existing contents and introd sections.

show more

Finalize version

Comments

New Comment

State All states

Sort per date Descending

No results

Create

Basic Metadata

Context management

Textual information/summary, method

Further metadata

Reference and relation

Change log

Version management

Change message (create, selection, retention, version, transformation, anonymisation)

Fig.3: Data file element page presenting fields for auto generated and manual metadata entries.

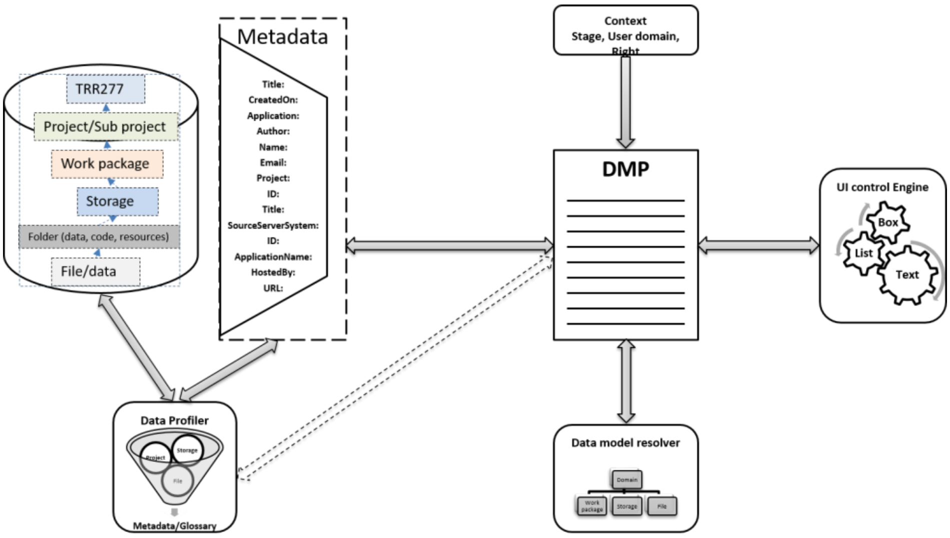


Fig.4: System design, components, and interaction model for dynamic DMP.

5.4.3. Data profiler

Profiler component will serve two purposes. On one hand it will iterate through the data and corresponding metadata to create a glossary, on the other hand it will fill the missing metadata fields of the corresponding data automatically or suggest users in case users attempt it manually. It will also serve as an alternate source of information for DMP component.

5.4.4. DMP

DMP component represents DMP form including its data model as well as its controls. DMP controls will be communicating with the metadata component of the associated data. If the required information is available on the data part, the

DMP will consume it or suggest the user for input. If the information is missing on the data part, then DMP user as well as data part, e.g., data manager, data system could be informed. Or in an ideal case data could be labelled with such a fact. In some cases, DMP control will also be communicating with profiler component, e.g., if a metadata field is empty then glossary could be consulted, or profiling function could be invoked on ad hoc bases etc. DMP component will also have an associated context component which will be governing contextual information and states of DMP.

5.4.5. Data model resolver

Since DMP can be associated with different levels of data, a data model resolver component will resolve DMP form, data model etc. as per corresponding dataset.

5.4.6. UI control engine

Since the values of the DMP are supposed to be dynamic, the user interface components should also be rendered dynamically as per corresponding data models, e.g., checkboxes and lists. Therefore, UI control engine will be resolving form and its components' rendering at the run time.

5.5. Methodology and Future Work

Based on the problem definition, we are following Design Science Research Methodology (Hevner et al., 2004) to produce and evaluate artefacts iteratively. With the inception of AMC in 2020, RDM policy and strategy, custom DMP and corresponding data lifecycle model have already been defined. First the solutions were evaluated and improved based on the feedback of test users who were researchers in the field of architecture. Then the solutions were offered to the community for their real jobs. Data management and organisation is being done using TUM Workbench as VRE (Shah, 2022). The work on data model based on custom DMP is in progress. After the completion of the data model the proposed system will be implemented. The produced artefacts will be offered to the AMC research community and the evaluation and improvement processes will be started iteratively.

6. CONCLUSION

This paper discussed two main interrelated topics. First topic about RDM strategies & systems and the second topic about system design for dynamic DMPs. The strategies and systems for RDM have been discussed and categorised based upon complexity and scale in an incremental way. In this way readers could be guided to define their own strategy and selection criteria based on their own use cases and requirements as well as resources. Second section has discussed advanced documentation system, i.e., DMP. The system design has been presented based on an ongoing work. It indicates as to how an integrated DMP could be realized which will be verifiable, semi-automatically updatable and maintainable for common understanding and to increase its acceptance. Further development of features, e.g., integration of data sources like Data Science Storage (DSS) in TUM Workbench are in progress. Although system design for dynamic DMP was presented on the example of TUM Workbench, the approach is system agnostic and could be adopted for any other RDM system exhibiting similar features. Members and the research work of AMC belongs to different disciplines. It would help to generalise and determine the limitations of the proposed system. Therefore, future work would include the discussion of results, challenges, lessons learnt, strength and limitation in handling various types of data based on the proposed strategy.

ACKNOWLEDGEMENTS

The work presented in this paper is being conducted as part of the collaborative research centre 'Additive Manufacturing in Construction - The Challenge of Large Scale' funded by the German Research Foundation (DFG), within the I - Information Infrastructure Project.

REFERENCES

- AMORIM, R. C. et al. 2017. A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society,* 16**,** 851-862.
- BAKER, M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature,* 533**,** 452-454.
- BORGHI, J. A. et al. 2018. Support Your Data: A Research Data Management Guide for Researchers. *Research Ideas and Outcomes,* 4.
- CANDELA, L., CASTELLI, D. & PAGANO, P. 2013. Virtual research environments: an overview and a research agenda. *Data science journal***,** GRDI-013.
- CLARIN-D. *Wizard for Data Management Plan Creation* [Online]. Available: https://www.clarin-d.net/en/preparation/data-management-plan [Accessed 07.07.2022].
- DATA CITATION SYNTHESIS GROUP. 2014. *Joint Declaration of Data Citation Principles.* [Online]. San Diego CA: FORCE11. Available: https://force11.org/info/joint-declaration-of-data-citation-principles-final/ [Accessed 07.07.2022].
- DCC. 2021. *Towards better efficiency – integrating data management plans with institutional systems* [Online]. Digital Curation Centre. Available: https://dcc.ac.uk/news/towards-better-efficiency-integrating-data-management-plans-institutional-systems [Accessed 07.07.2022].
- DFG. 2021. *Handling of research data* [Online]. DFG. Available: https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/forschungsdaten_checkliste_en.pdf [Accessed 07.07.2022].
- DORA. 2012. *San Francisco Declaration on Research Assessment (DORA)* [Online]. Available: https://sfdora.org/ [Accessed 07.07.2022].
- EASTERBROOK, S. M. 2014. Open code for open science? *Nature Geoscience,* 7**,** 779-781.
- EC H2020. 2018. *TEMPLATE HORIZON 2020 DATA MANAGEMENT PLAN (DMP)* [Online]. European Union. Available: https://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotated_en.pdf [Accessed 07.07.2022].
- EUROPEAN COMMISSION. *Data management* [Online]. European Union. Available: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm [Accessed 07.07.2022].
- FOSTER. *What is Open Science? Introduction* [Online]. FOSTER. Available: https://www.fosteropenscience.eu/content/what-open-science-introduction [Accessed 07.07.2022].
- FREUDENBERG, M. et al. The metadata ecosystem of DataID. Research Conference on Metadata and Semantics Research, 2016. Springer, 317-332.
- GIT-SCM. *Getting Started - About Version Control* [Online]. Available: https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control [Accessed 07.07.2022].
- GRAY, J. et al. 2005. Scientific data management in the coming decade. *Acm Sigmod Record,* 34**,** 34-41.
- HEVNER, A. R., MARCH, S. T., PARK, J. & RAM, S. 2004. Design science in information systems research. *MIS quarterly***,** 75-105.
- HICKS, D., WOUTERS, P., WALTMAN, L., DE RIJCKE, S. & RAFOLS, I. 2015. Bibliometrics: The Leiden Manifesto for research metrics. *Nature,* 520**,** 429-431.
- JONES, S., WHYTE, A. & PRYOR, G. How to Develop Research Data Management Services - a guide for HEIs. 2013. Digital Curation Centre (DCC).
- KLEIN, S. 1993. Information Logistics. *Electron. Mark.,* 3**,** 11-12.
- LEEK, J. T. & PENG, R. D. 2015. Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences,* 112**,** 1645-1646.

- MARCIO, T. M., CHRISTINE, O. & OMG. 2017. *Hybrid Cloud Considerations for Big Data and Analytics* [Online]. Cloud Standards Customer Council. Available: https://www.omg.org/cloud/deliverables/hybrid-cloud-considerations-for-big-data-and-analytics.htm [Accessed 07.07.2022].
- MESIROV, J. P. 2010. Accessible Reproducible Research. *Science,* 327**,** 415-416.
- MIKSA, T., OBLASSER, S. & RAUBER, A. 2021a. Automating Research Data Management Using Machine-Actionable Data Management Plans. *ACM Trans. Manage. Inf. Syst.,* 13**,** Article 18.
- MIKSA, T., RAUBER, A., GANGULY, R. & BUDRONI, P. 2017. Information Integration for Machine Actionable Data Management Plans. *Int. J. Digit. Curation,* 12**,** 22-35.
- MIKSA, T., SIMMS, S., MIETCHEN, D. & JONES, S. 2019. Ten principles for machine-actionable data management plans. *PLOS Computational Biology,* 15**,** e1006750.
- MIKSA, T. et al. 2021b. Application Profile for Machine-Actionable Data Management Plans.
- OPENAIRE. 2017. *What is a Data Management Plan (DMP) and how do I create one?* [Online]. Available: https://www.openaire.eu/what-isa-data-management-plan-and-how-do-i-create-one [Accessed 07.07.2022].
- OPENAIRE AND EUDAT. *ARGOS* [Online]. Available: https://argos.openaire.eu/splash/ [Accessed 07.07.2022].
- RDMO. *Research Data Management Organiser (RDMO)* [Online]. Available: https://rdmorganiser.github.io/ [Accessed 07.07.2022].
- SHAH, S. A. H. 2022. Research data management of large scale projects and a reference model of data life cycle for dynamic DMPs. Forum Bauinformatik, 2022. 157-165.
- SMALE, N., UNSWORTH, K., DENYER, G., MAGATOVA, E. & BARR, D. 2020. A Review of the History, Advocacy and Efficacy of Data Management Plans. *International Journal of Digital Curation,* 15**,** 1-29.
- SPRECKELSEN, F., RÜCHARDT, B., LEBERT, J., LUTHER, S., PARLITZ, U. & SCHLEMMER, A. 2020. Guidelines for a Standardized Filesystem Layout for Scientific Data. *Data,* 5**,** 43.
- STODDEN, V. The Scientific Method in Practice: Reproducibility in the Computational Sciences. 2010.
- STODDEN, V., GUO, P. & MA, Z. 2013. Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. *PLOS ONE,* 8**,** e67111.
- TENOPIR, C., ALLARD, S., DOUGLASS, K., AYDINOGLU, A. U., WU, L., READ, E., MANOFF, M. & FRAME, M. 2011. Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE,* 6**,** e21101.
- THE ROYAL SOCIETY 2012. Science as an open enterprise. The Royal Society.
- VAN GORP, P. & MAZANEK, S. 2011. SHARE: a web portal for creating and sharing executable research papers. *Procedia Computer Science,* 4**,** 589-597.
- WEBER, T. & KRANZLMÜLLER, D. 2019. Methods to Evaluate Lifecycle Models for Research Data Management. Available: https://ui.adsabs.harvard.edu/abs/2019arXiv190111267W [Accessed January 01, 2019].
- WILKINSON, M. D. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3**,** 160018.
- WILSON, G. et al. 2014. Best Practices for Scientific Computing. *PLOS Biology,* 12**,** e1001745.