ARPHA Preprints

PENSOFT

Data Management Plan

Author-formatted document posted on 24/02/2023

Published in a RIO article collection by decision of the collection editors.

DOI: https://doi.org/10.3897/arphapreprints.e102609

Deliverable D12.9 Data Management Plan

Lyubomir Penev, Teodor Georgiev, Boris Barov, Pavel Stoev, Kristina Hristova

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102609

Data Management Plan

Deliverable D12.9

29 October 2021

Lyubomir Penev, Teodor Georgiev, Boris Barov, Pavel Stoev, Kristina Hristova

Pensoft Publishers Ltd.

BiCIKL

BIODIVERSITY COMMUNITY INTEGRATED KNOWLEDGE LIBRARY

This project receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492.

Start of the project:	May 2021
Duration:	36 months
Project coordinator:	Prof. Lyubomir Penev Pensoft Publishers
Deliverable title:	Data Management Plan
Deliverable n°:	D12.9
Nature of the deliverable:	ORDP: Open Research Data Pilot
Dissemination level:	Public
WP responsible:	WP12
Lead beneficiary:	Pensoft Publishers
Citation:	Penev, L., Georgiev, T., Barov, B., Stoev, P. & Hristova, K. (2021). *Data Management Plan*. Deliverable D12.9 EU Horizon 2020 BiCIKL Project, Grant Agreement No 101007492.
Due date of deliverable:	Month 6
Actual submission date:	29 October 2021

Deliverable status:

Version	Status	Date	Author(s)
1.0	Draft	18 October 2021	Pensoft
2.0	Review	27 October 2021	MeiseBG, LifeWatch, GBIF, Plazi
3.0	Submission	29 October 2021	Pensoft

Table of contents

Preface

The main goal of the BiCIKL project is to improve, for the first time, seamless access, linking and usage tracking of data within a network of Research Infrastructures managing different data classes (literature, specimens, samples, occurrences, sequences, taxon names and Operational Taxonomic Units (OTU)), ultimately represented also in a biodiversity knowledge graph. To achieve this, the consortium members will operate with huge amount of data during and after the end of the project.

As a Horizon 2020 project, BiCIKL conforms to the Open Research Data Pilot (ORDP)[1] and Article 29.3 of the H2020 Model Grant Agreement by default, hence the consortium aims to improve and maximise access, sharing, linking and reuse of FAIR Open Research Data (ORD), generated or managed by the project. A detailed Data Management Plan is a critical part of the ORDP. The DMP described in the present document is developed in BiCIKL within the first six months of the project and it will evolve as a "*living document*" during the lifetime of the project and beyond in order to present the status of the project's reflections on data management.

The BiCIKL DMP outlines the handling of research data and provides the basis of the project consortium's data management life cycle for the data collected, generated and processed by the participants in the project. The DMP also covers the methodologies and standards previously developed for data sharing and open access, curation and preservation. The subject of the DMP is the management of research data. Personal data management is covered by deliverable *D9.1 Protection of Personal Data*.

The BiCIKL DMP was developed in close collaboration with all project partners and involved Research Infrastructures (RI) who provided information on their data management practices and policies in a questionnaire and planned generation, collection, and processing of data for the purposes of building a resilient data management strategy of the project which meets all criteria for open research.

This DMP aims to adhere to the FAIR (Findable, Accessible, Interoperable, Reusable) data management criteria of Horizon 2020[2].

Summary

The BiCIKL DMP is structured in five sections, which aim to establish the scope and terms of use of research data within the project in accordance with the Horizon 2020 requirements of data management.

The first section provides an introduction to the plan, which outlines the main types of data and the management practices that BiCIKL implements throughout the three-year project

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

[2] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

duration, as well as aspects of sustainable management of results and data after the conclusion of the project.

The second section of the document provides an overview of the commitments that BiCIKL has made in relation to handling data in a controlled and transparent way, and ensuring open access to research data and results in line with the EU's Open Research Data Pilot and FAIR data management.

The third section describes the details of data management within the project, focusing on different aspects of the process - from data collection, through processing, to storage and access provision.

The fourth section features the main obstacles and risks, identified by the partners and recommendations for possible solutions. This part covers risks from technical and social points of view. Taking proper measures is of importance, as BiCIKL project focuses on the linking of biodiversity data which would require different types of research data, and therefore different specific data management practices to be added to the general data management framework.

The fifth section of the DMP features concluding remarks on the data management strategy adopted by the project, and it outlines future updates and additions to the plan, which are going to be presented at a later stage of the project's development.

List of abbreviations

API	Application Programming Interface
DMP	Data Management Plan
EU	European Union
FAIR	Findable, Accessible, Interoperable and Re-usable
GDPR	General Data Protection Regulation
ORDP	Open Research Data Pilot
POPD	Protection of Personal Data
RI	Research Infrastructure

1. Introduction

As part of WP12 Project Management, BiCIKL has developed a detailed plan for sustainable management of research data used within the project. The current document is based on the current data management practices provided by the BiCIKL partners, and specifies the foundations of the management life cycle of the data collected, generated and processed by the project partners and their research infrastructures. The BiCIKL DMP addresses all key aspects of data management, including the handling of research data during and after the end of the project, data collection and data processing by the project, as well as data publishing, storage and preservation.

Numerous and varying datasets will be collected and generated by the project partners and involved research infrastructures in the framework of BiCIKL. The project aims to not only openly share data but to provide a unique new level of linking open data that will open the full research cycle and will encourage reuse and reproducibility along the line: specimens ➔sequences ➔ species ➔ analytics ➔ publications ➔ biodiversity knowledge graph ➔ re-use.

To ensure interoperability, the BiCIKL project aims to collect and document the data in standardized formats (i.e. Darwin Core) to ensure that the datasets can be understood, interpreted and shared with accompanying metadata and documentation and relevant supporting material. Metadata standards will depend on the discipline and/or the methodology used to produce the data. BiCIKL partners use discipline-specific repositories and common/standard metadata requirements and ontologies, e.g. TaxPub, OpenBiodiv-O, Ecological Metadata Language (EML), the latter being a metadata specification particularly developed for ecology and biodiversity science, and where necessary append other existing ISO-90155 compliant metadata libraries dependent on discipline-specific or institutional repositories.

BiCIKL's main focus is on data that are already open or such that will be made open and FAIR during the project and beyond it. The ambitious objectives and the implicit interdisciplinarity of the project's research tasks include a variety of research methods. This results in diverse research data, and makes the importance of a well-developed data management strategy from the beginning of the project even higher.

2. Open access statement and data sharing

In the context of the Horizon 2020 programme, the European Commission launched the Open Research Data Pilot (ORDP) which aims to ensure that all research data generated within the Horizon 2020 research projects are findable, accessible, interoperable, and reusable (FAIR). The primary purpose of the FAIR data prescription is to maximise the utility of research data by ensuring open access to scientists. The Horizon 2020 Open Research Data Pilot is the most comprehensive EU mandate for open access to scientific publications, research data, and metadata produced with Horizon 2020 funding.

The DMP ensures that the research activities of the BiCIKL project are compliant with the Horizon 2020 Open Access policy and the recommendations of the ORDP, in all cases in which these apply. In this context, the BiCIKL DMP outlines how research data will be collected, processed or generated within the project and how these data will be curated and

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102609

preserved during and after the project, in line with the main goal of the project: **improved and increased access to and use of interlinked and FAIR open biodiversity data**.

FAIR data are findable, accessible, interoperable and reusable, and applying these principles to the BiCIKL project aims to ensure data are robustly managed. The BiCIKL DMP thus features key initial provisions on ensuring that the FAIR data guidelines are followed, in as much as it is humanly and machine possible, and the project's research data are managed in accordance with Horizon 2020 prescriptions for ensuring the open access and reuse possibilities of data generated and processed within the project.

In a separate section 1.2.4. of the project description, BiCIKL declares itself as "*an entirely open science project, from start to end*". The open science agenda assumes several more aspects than opening and sharing of research data and can be summarised in the following **BiCIKL Open Science Statement**:

BiCIKL aims to conform whenever possible to all fundamental principles of open sciences: open access, open data, open source, human- and machine-readability, interoperability, transparency and reproducibility of the full research cycle.

3. Data overview and management

The BiCIKL project will use, generate and link a wide variety of datasets from different fields of research and access them through multiple data gathering methods. The project partners will generate and process both primary and secondary research data. Types of data are going to be collected from various sources, including publicly available datasets, institutional or private data, as well as self-generated data. Some of the RIs involved in BiCIKL allow their registered users to develop their own datasets, e.g. in PlutoF, Zenodo, GBIF and others.

The participating Research Infrastructures provide access to data, associated tools and services at each separate stage of, and along the entire research cycle. BiCIKL will specifically focus on providing new methods and workflows for an integrated access to harvesting, liberating, linking, accessing and re-using of subarticle-level data (specimens, material citations, samples, sequences, taxonomic names, taxonomic treatments, figures, tables) extracted from literature.

The data provided below are gathered from all BiCIKL partners by completion of a Data Management Questionnaire. The results indicated below are expressed as percentages of the total number of partners.

3.1. Origin of data

The following categories of data origin stand out as more frequently used by partners (in % of respondents using each data type, categories may overlap):

Publicly available datasets	81.25%
Institutional or private data (copyrighted; permit required)	43.75%
Self-generated data	43.75%

Other	12.50%

Some partners specified that registered users are free to develop their datasets in the Research Infrastructures that they are representing, e.g. PlutoF has 7,000+ registered users with their own datasets and Zenodo also uses datasets from researchers world-wide.

Recommendation: *BiCIKL will ensure that the origin and provenance of the data are openly recorded and displayed where appropriate, especially at metadata level, through citation or provenance records. Such practice will allow appropriate citation and credit to data collectors, managers and owners. The requirement for attribution is implied by several use licenses (e.g. Creative Commons CC-BY), however, it is also a long-established and community-agreed element of the research ethics.*

3.2. Acquisition of data and reuse permission

The research data acquired by project partners are registered under different use licences, including:

Creative Commons CC0 Public Domain Dedication	68.75%
Creative Commons Attribution (CC-BY) License	37.50%
Open Data Commons Attribution License	31.25%
Open Data Commons Public Domain Dedication and License	6.25%
Other	56.25%

Most partners use two or more licenses.

Recommendation: *In terms of acquiring reuse permission for data that is property of an external research institution, business organisation, or another type of legal entity, project partners are responsible for obtaining a licence as a method to secure reuse permission.*

3.3. Data collection

Data are collected for the purposes of the research tasks performed within the project's lifetime. Most partners implement a combination of two or more data collection procedures. The main data collection methods used, according to the preliminary survey among project partners, are as follows:

Natural history collections	56.25%
Literature review & harvesting	50.00%
Field work	31.25%
Sequencing	31.25%

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102609

Data submission/publishing by third parties	2.50%
Interview surveys	0.00%
Other	0.00%

Recommendation: *The methods of data collection should be explicitly recorded in the metadata, following community accepted standards (e.g. BasisOfRecord in Darwin Core).*

3.4. Type of data collected

The categories of data which are collected by the partners and research infrastructures, participating in BiCIKL are:

Taxon names	87.50%
Persons	87.50%
Specimens	75.00%
Molecular sequences	62.50%
Articles	56.25%
Images	50.00%
Taxon treatments	37.50%
Other	50.00%

Besides these key data types that are of primary interest to BiCIKL, the respondents mentioned that they also collect gene and gene products, variants, diseases, gene functions, drugs and chemicals, other specimen metadata (e.g. location, date) and specimen images, material citations, taxon hypotheses, functional trait data, data from molecular and other laboratories, collections, institutions, facilities and instruments, references, possibly type specimens, specimens = materials citations, sampling data and sample processing data.

Recommendation: *The ultimate goal of BiCIKL is to provide linking mechanisms between different data types, and access to interlinked data. Hence, partners must ensure that each data type they collect and manage is made FAIR to allow seamless access, linking and re-use. In addition, the project will elaborate standards and guidelines for linking between different data classes, to ensure interoperability between sub-domains in the realm of biodiversity.*

3.5. Use of identifiers

The following types of identifiers are associated with the data managed in BiCIKL:

Globally unique IDs, resolvable	87.50%

Local, not globally unique	50.00%
Globally unique, non-resolvable	37.50%
Other	18.75%

Recommendation: *The persistent identifiers (PIDs) of data and research objects make the primary focus of BiCIKL. The project builds upon the already existing structures and recommendations (e.g. CETAF Identifiers Guidelines[3]) or long used accession numbers for genetic sequences used by the INSDC consortium. BiCIKL will carry out user requirements surveys, discussion and hackathons to produce guidelines and recommendations for types of PIDs to be used for the different data types, including innovative solutions for sub-article level data in the published literature (e.g., treatments, material citations, accession numbers, images, tables, and others). The ultimate goal of the project is to build a sound basis for norms and practices in use of PIDs in biodiversity data management, research and publishing.*

3.6. Data storage

The most common formats in which data are preserved and exported are:

XML	62.50%
CSV	43.75%
TSV	43.75%
RDF	37.50%
Other	75.00%

Other commonly used formats are JATS, JSON, DwC Archive, JSON. Partners and research infrastructures operate with different sizes of data - from 100GB up to 20PB.

Most of the BiCIKL partners - 75% use their own infrastructure to store data. Typically, server locations are within the EU, by 68.75%.

Recommendation: *We recommend that partners choose data formats for final data deposition which are lightweight, non-proprietary and widely used (e.g. CSV, XML, JSON, RDF etc.). Partners should take all the necessary steps to protect their datasets and run regular backups on two independent servers at least.*

3.7. Ownership of data management platforms

The data management platforms used by most BICIKL partners and research infrastructures are distributed according to their ownership and accessibility as follows:

[3] https://cetafidentifiers.biowikifarm.net/wiki/

ARPHA Preprints *Author-formatted document posted on 24/02/2023.* DOI: https://doi.org/10.3897/arphapreprints.e102609

Open source	68.75%
Proprietary own	12.50%
Proprietary licensed	12.50%

Recommendation: *We cannot request all partners to put in open source their background data management platforms, however all software tools and workflows produced by the project should be available as open source projects.*

3.8. Software frameworks

The software frameworks used as principal data management platforms by the project partners are:

- Custom type
- PHP
- MySQL
- Java
- MongoDB
- Lucene
- BG-Base (designed and licensed by an external company specifically for Botanic Gardens)
- SOLR search engine
- PostgreSQL
- Elasticsearch
- Django REST framework
- Ember.js framework
- Invenio: http://inveniosoftware.org
- Collection management systems
- CORDRA
- European Loans and visits system
- CETAF register
- GraphDB (proprietary licensed)
- Cloudera Hadoop
- Oracle
- Vertica
- Compute clusters including Kubernetes)
- Web servers and websites with a set of RESTful interfaces.

Recommendation: *The great variety of software platforms, tools and frameworks requires the elaboration of common standards for access to and linking between data to both humans and machines. These standards will be agreed by all partners and put in place for adoption to the future members of the BiCIKL community.*

3.9. Application Programming Interfaces (APIs)

The majority of project partners (75%) provide an Application Programming Interface (API) access to their data. The most common type of APIs that BiCKL researchers use are:

- FUB-BGBM: REST API as well as SPARQL endpoint for the Botany Pilot
- GBIF API: https://www.gbif.org/developer/summary
    - The web portal https://www.botanicalcollections.be has a REST API, which is used by its front-end, but it is currently not documented. There is also an RDF/XML endpoint for individual specimen data, following the CETAF identifier principles: https://cetafidentifiers.biowikifarm.net/wiki/ ."
- SIB/SIBiLS and University of Tartu/PlutoF: REST API-s.
- CERN/Zenodo: API documentation: http://developers.zenodo.org
- LifeWatch provide for metadata the following APIs: https://metadatacatalogue.lifewatch.eu/doc/api/index.html
- DiSSCo: REST and DOIP through CORDRA instances, documentation: https://www.cordra.org/documentation/api/index.html
- COL has a JSON REST API: https://api.catalogueoflife.org
- MNHN: API & Regex to retrieve links
- OpenBiodiv: SPARQL endpoint available; no REST/GET or GraphQL API
- EMBL ENA: APIs are described here:
    - data access: https://ena-docs.readthedocs.io/en/latest/retrieval/programmatic-access.html
    - data submission: https://ena-docs.readthedocs.io/en/latest/submit/general-guide.html
- TreatmentBank provides an API a fine-grained query API on top of its data statistics (https://tb.plazi.org/GgServer/dioStats and https://tb.plazi.org/GgServer/srsStats), supporting JSON, XML, and tabular formats, and a coarse API for its XML treatments (https://tb.plazi.org/GgServer/search)
- RefBank provides an XML based API (https://tb.plazi.org/RefBank/static/downloadRefBank.html)

The most widely spread kind of API access is REST - 68.75% of the cases.

Recommendation: *APIs are the gate to data access and sharing, hence their elaboration and improvement are the* ***ultimate task for all partners and RIs partnering in BiCIKL****! APIs can be of different types (e.g. REST, GraphQL) and their implementations have constraints and advantages. The BiCIKL RI Technical Forum will discuss the different options and provide guidelines ensuring machine-actionable access to and bilateral linking of data via APIs. Partners will have to optimise their APIs to ensure successful fulfilment of BiCIKL's mission, tasks and goals.*

3.10. Data ownership

In the general case data are owned by the submitters of data of their research infrastructures. However, some exceptions exist and the following partners: BGBM, MeiseBG, Pensoft and CEFAF state that they are the formal owners of data.

Recommendation: *According to common understanding, both in the EU and worldwide, data itself is not a creative product, hence it cannot be a subject of intellectual property rights (mostly copyright), except for substantial databases (EU law). 'Ownership' on data may have different forms (e.g. data hoarding), however BiCIKL will minimise use of data which are not openly available at appropriate liberal open licenses; data produced by the project will be open by default.*

3.11. Data management and use in the context of BiCIKL

BiCIKL data are openly shared through automated workflows with relevant repositories, including but not limited to the Biodiversity Literature Repository at Zenodo, Plazi TreatmentBank, GBIF, ENA, CoL, OpenBiodiv, PMC Europe.

All data and models, both generated as part of BiCIKL and obtained from other sources, will be annotated, using internationally recognised keywords and meta-tags. Outputs from BiCIKL will be organised in an easily accessible and interpretable format. The necessary tools, standards and protocols for making BiCIKL data accessible, findable, exchangeable and secured in the long term will be made available by the BiCIKL partners to the BiCIKL users.

Most of the BiCIKL partners and involved RIs will not need to change the licence types they are using for providing their regular services, and the data provided from them in the timeframe of BiCIKL and beyond will be entirely open. However, some partners mentioned that selected data may be embargoed for a limited time period during ongoing research projects. Sensitive specimen data may also be accessed in different geographical or other types of scales. Specimens flagged with serious data quality issues will be withheld until these issues are addressed.

The BiCIKL partners and RIs provide their data for the purposes of the project through the following ways:

Free access, no permission needed	93.75%
Contacting the institution/data owner	18.75%

Recommendation: *Additional permission from third-parties for data used in BiCIKL will not be needed as long as the data are open under appropriate licenses. If submitters decide to close access, then this will require permission to access.*

At this stage of the project some of the partners need to change their data management structure to make their data FAIR. The following types of issues were identified among the given share of the BiCIKL partners:

Bilateral linking with data from the participating RIs	62.50%
Search, discovery and access mechanisms	37.50%
Resolvable persistent identifiers	31.25%
API in accordance with the project specifications	31.25%
Appropriate use licenses and policies	18.75%

Other aspects of data management	18.75%

3.12. Use of metadata

Regarding the provision of meta-data, 87.5% of the partners provide metadata and the most commonly used metadata type and standards:

- Darwin Core
- TaxPub
- ABCD
- ELN
- JATS
- EML
- CETAF Specimen Preview Profile
- Dublin Core
- GSC
- MCL
- BibTeX
- CSL
- DataCite
- Dublin Core
- DCAT
- JSON
- JSON-LD
- GeoJSON
- MARCXML
- LifeWatch ERIC Metadata Catalogue
- ODS (openDigital Specimen) w
- ABCD
- AudubonCore
- MODS

Recommendation: *The variety of metadata standards and formats should not prevent partners from making their metadata FAIR whenever needed. This process should be ensured through various measures at technical, policy and management levels.*

3.13. Personal data

Some of the partners collect personal data in order to provide their services, e.g. contact persons (creator, metadata provider, contact points, associated parties) by storing name, surname, email address. Therefore the necessary personal data protection measures are of importance for the project.

Use of personal data in research datasets requires a necessary amount of anonymisation. Therefore, data preparation in the project necessarily includes technical steps for achieving a required level of anonymisation of entries.

Personal information for non-research purposes is regulated in accordance with the EU requirements of the GDPR, with each institutional partner that would collect and store personal data providing the necessary protection of the data, and performing administrator roles that ensure personal data are not only secure, but also that they are not used for purposes unrelated to the project.

Each institutional partner has an appointed data protection officer (DPO), and has presented their contact information to the consortium members and the European Commission in Deliverable D9.1 POPD.

Recommendation: *Use of personal data is a strictly regulated topic, and expressed consent from the parties that provide their data voluntarily is required for the purposes of data collection within the project. We implement the provision of consent forms translated into relevant languages upon requesting and/or receiving research-relevant data from individuals and/or organisations.*

3.14. Sensitive data

37.50% of the partners collect sensitive data like data about protected and invasive alien species.

Recommendation: *The appropriate handling of sensitive data is a responsibility of each institution and the research teams who gather the data.*

3.15. Data publishing

The majority of partners (87.50%) publish the data used in or related to BiCIKL through their own or external publishing platforms, as follows:

Own publishing platform	62.50%
GBIF	56.25%
Zenodo	43.75%
Data papers in journals	37.50%
Dryad	6.25%
Figshare	0.00%
DataONE	0.00%
Other	12.50%

Recommendation: *BiCIKL follows the guidelines on open access to scientific publication and research data stated in Horizon 2020. Deliverables and other important project outputs will be published in a dedicated open access collection in RIO Journal.*

Data publication should ensure a PID of the published dataset, citation mechanism, dissemination and usage tracking.

BiCIKL partners will distinguish between the (1) standalone data publishing (e.g. publishing of a dataset in GBIF or sequences in ENA) and (2) scholarly data publishing (e.g. data published in academic journals in supplementary files, as structured data in the article narratives or in the form of data papers).

Furthermore, the software tools or plugins produced within the BiCIKL will be available as open source code under an appropriate license and published in the open science RIO Journal to ensure findability and reusability of all open source resources.

Scientific results will be freely disseminated through appropriate channels including scientific publications, presentations at international conferences and workshops. The publication venues will be primary scientific high-impact open access journals (such as the European Journal of Taxonomy (EJT), Biodiversity Data Journal (BDJ), Research Ideas and Outcomes (RIO), or others.

4. Potential RISKs and possible solutions

Some potential RISKs and risks that could prevent delivering FAIR data and achieving its interoperability from technical and social perspectives were identified:

- Importing biodiversity literature from closed Access content is challenging.
- Technical limitations of the botanicalcollections.be web portal and/or GBIF, such as the scalability of the RDF endpoint.
- Maturity of the DiSSCo infrastructure and the openDS standard.
- Technical infrastructure to serve data directly as FAIR digital objects might not be yet in place and may not be in place before the end of BiCIKL. Lack of appropriate actions and initiative from certain RIs too.
- Human resources, financial resources, technical maturity of the institutions and society.
- FAIR technical offerings might not be compatible with established open data protocols and conventions (e.g. INSDC ID systems).
- Independent publishers might not have technical abilities to retrieve the data.
- Very long term storage (beyond CERN life time) might be an issue.
- Lack of interoperability between standards used by different communities. Even if the same standards are used, data are often published as free text rather than linking to semantic resources.
- License incompatibility and limitations in data standards, in particular Darwin Core.
- Lack of resolvable persistent identifiers, heterogeneity in collection management systems and lack of infrastructure for annotations and linkages, but there are also potential obstacles in social (social change will be needed to support digital (transformation) and policies (alignment).

The risks identified above can be attributed to various reasons, some of which are beyond the limit of control of BiCIKL. Those that can be managed and mitigated within the project will be addressed as follows:

- The technical obstacles in the implementation of the access and linking mechanisms will be in the focus of discussions within the RI Technical Forum held monthly. The Forum will identify the technical problems in access and linkage the RI face and

provide recommendations and advice. The Forum will also ensure that the recommended standards for use of PIDs and APIs will be thoroughly discussed and agreed upon by all RIs before taking these into action.

- Special attention will be paid to the risk of choosing methodologies that are not suitable for the purpose and additionally may be not accessible to or appropriate for all RIs and data. A good example for that is the API type the different RIs may decide to use; in several cases a REST API will be sufficient, in other GraphQL will be more appropriate and in the third case a SPARQL endpoint would be the only solution. BiCIKL will follow an agile and flexible approach in such cases, in the understanding that there is virtually no way one solution can fit all. Compromise and multiple solutions will be sort to achieve the main goals of the project.
- Technical obstacles might be combined with a misunderstanding of the obligations in providing symmetrical access to linked data, therefore bilateral and multilateral agreement templates will be provided to encourage the partners and RIs to pursue long-term agreement for provision and maintenance of services.

5. Conclusions

The BiCIKL DMP aims to provide a framework for the governance of data within the project. In order to create the framework, project partners must coordinate their data management practices and establish the basic rules for data collection, processing, and protection within the project.

As the DMP is created at an early stage of the project development, the Project Management Team will take the necessary steps to keep the document updated.

The recommendations made in the current DMP will be exported, synthesised and designed into BiCIKL Data Management Guidelines, which will be circulated among the project partners. The purpose of these guidelines will be to raise awareness and streamline the approach to data management, as well as to provide concise and useful information on data management.