H2020-INFRADEV-2019-2
Grant Agreement No 871043

DATA MANAGEMENT PLAN

DiSSCo Prepare WP9 - Task 9.2 – Deliverable 9.2

WP Leader: Naturalis Biodiversity Center
Author(s): Dimitris Koureas, Eva Alonso

Deliverable references

Deliverable	D 9.2 DATA MANAGEMENT PLAN
Work Package	9
Lead Partner	NATURALIS
Status	DRAFT
Deliverable type	REPORT
Dissemination level	PUBLIC
Due date	30/04/2020
Submission date	29/04/2020
DOI	https://doi.org/10.34960/8mzj-g791

Abstract

The DiSSCo Prepare Project Data Management Plan (DMP) provides guidance to the beneficiaries of the Project in regards to practices in generation, collection and stewardship of project data. The DMP is taking into account the FAIR principles and fully supports open science. All data and software code generated by DiSSCo Prepare Project will, by default, become publicly accessible for all, but commercial, purposes. The DMP provides guidance on the recommended licenses to be assigned to all project outputs by their joint owners. Finally the DMP acknowledges the need for all outputs to become readily available for re-use by the future legal entity of the DiSSCo RI.

Keywords

FAIR, Data Management Plan, DiSSCo, Open source, Open access

H2020-INFRADEV-2019-2
Grant Agreement No 871043

INDEX

INTRODUCTION

DiSSCo Research Infrastructure (RI) has entered the European roadmap for RIs in 2018 when its Preparatory Phase has started. During this phase, DiSSCo aims at raising its overall maturity and position itself better to implement its elaborate implementation programme. The DiSSCo Prepare project is the primary vehicle through which DiSSCo RI is improving its readiness, in view of a more elaborate implementation phase.

Specifically, DiSSCo Prepare project has two primary goals:

- Improve DiSSCo's overall Implementation Readiness Level (IRL), that is, its ability to execute construction and effect related actions based on clear, actionable guidelines with minimum risk. DiSSCo Prepare will raise DiSSCo's IRL across five critical dimensions: scientific, technical, organizational, financial and data readiness;
- Deliver the DiSSCo Construction Masterplan.

In the context above, the project aims at generating knowledge and data directly applicable to the implementation phase of the DiSSCo RI, with the forthcoming DiSSCo legal entity and the DiSSCo national nodes being the entities that will benefit the most out of work carried out in the project.

The current Data Management Plan is developed, considering the following:

1. The specificities of the project, in relation to the broader preparatory phase programme (portfolio of multiple projects) and the pre-identified user of all outputs (forthcoming DiSSCo RI legal entity);
2. The provided by the European Commission template for developing DMPs;
3. The planned development of the DiSSCo knowledge base and the incorporation of all knowledge outputs of DiSSCo-linked projects into a coherent corpus of knowledge;
4. The participation of DiSSCo Prepare to the EC open access pilot;
5. The commitment of DiSSCo RI and the DiSSCo community to the principles of open science;
6. The relevant articles in the DiSSCo Prepare Consortium Agreement regarding ownership and licensing of project products;
7. The continuous developments of the socio-technical framework for open and FAIR data.

The DiSSCo Prepare DMP should be considered a living document, which will be updated to reflect the latest developments around the project's data management throughout. The foundational aspects of the DMP, however, shall not change. These include the commitment of the consortium to the principles of open and FAIR data and open science practices.

Updated versions of this document shall be made available to the project's consortium and publicly and will replace any previous iteration of the DiSSCo Prepare DMP.

DATA SUMMARY

What is the purpose of the data collection/generation and its relation to the objectives of the project?

DiSSCo Prepare work programme includes a portfolio of activities that rely on the ability to collect, process and re-purpose information gathered. As the main preparatory phase project of DiSSCo RI, it uses this information as background knowledge to investigate the feasibility and applicability of different scenarios towards the implementation of the RI. The Project is concerned with areas of activity across five different dimensions (Table 1), and as such, the corresponding project beneficiaries will collect across a wide spectrum of data sources.

Table 1. Required types of data collected/generated across the IRL dimensions of the project and the anticipated linked data sources.

READINESS LEVEL	REQUIRED TYPE OF DATA COLLECTED/GENERATED (NON-EXHAUSTIVE LIST)	DATA SOURCES
SCIENTIFIC	• Use cases and use case analyses • RI User questionnaires • Other RI stakeholders questionnaires • Scientific publications and linked datasets • Natural Science collection descriptions datasets	• Other DiSSCo-linked project outputs (incl. ICEDIG, SYNTHESYS+, Mobilise Action) • Online tools for questionnaires • Open Access publications • Data provided by Natural Science Collections • Experts' opinions • Scientific publications • Natural History Museums data portals • Thematic international data aggregators • Dataset repositories • Outputs of other relevant projects • Website retrieved information • Survey results • Workshop and other meetings notes and minutes • DiSSCo facilities standard operating procedures / policies • CETAF previously collected data • DiSSCo Coordination and Support Office previously collected data
DATA	• Institutional policy documents • Training material (incl. audio-visual material and training datasets) • Survey-based collected data • Scientific publications and linked datasets	
ORGANISATIONAL	• Legal data • Targeted consultation-based data • User stories	
FINANCIAL	• Financial data/statements • Business models and business case analyses • Cost models • Survey-based collected data	
TECHNICAL	• Reference datasets (incl. specimen and specimen-derived information) • Software source code • Software technical documentation	

What types and formats of data will the project generate/collect?

Table 2. Data captured in relation to the project objectives, along with their expected format, origin and size (order of magnitude).

Data collected/generated	Project objectives	Formats	Origin	Expected (dataset) size (order of magnitude)
Data captured through user questionnaires - Data can include user preferences and other non-personal information collected	• To construct a service development framework focused on users in natural science collections-related research and research application (Task 1.1, Task 1.2) • To identify criteria for establishing a priority for the digitisation of natural science collections (Task 1.3) • To design Helpdesk and user support services that will provide the necessary information on the use of the infrastructure (Task 2.2).	Mostly text tabular formats (e.g. CSV)	Generated by DiSSCo Prepare tasks and collected through ICEDIG and other project outputs	10 MB
Data harvested through literature	All Project objectives	various	Collected through open access journals	100MB
Institutional policy documents	• Collate, refine and implement best practices for data mobilisation at the institutional level to develop the DiSSCo plan for data mobilisation and curation pipelines;	Mostly PDF and Doc(x) formats	Natural science collections and other institutions	100 MB
Financial data, including HR and other operational costs	Deliver DiSSCo cost book	Mostly in spreadsheet formats	Natural science collections and other institutions	100 MB
DiSSCo Prepare output documents metadata	All Project objectives	XML/JSON-LD	Project beneficiaries	100 MB

DATA UTILITY

Internal stakeholders

DiSSCo Prepare is focussing predominantly on delivering outputs relevant to the improvement of the overall Implementation Readiness Level of the DiSSCo RI. As such, it is expected that the main addressee and subsequent user, of all the project's outputs, is the legal entity (to be formed) that will represent the DiSSCo RI. The governance and management structures of the infrastructure shall be able to freely (re-)use any of the results of this project in any way relevant to the mission of the infrastructure. Additionally, the extended DiSSCo network of facilities and DiSSCo national nodes are expected to make use of many of the project outputs.

External stakeholders

Natural science museums and collections data infrastructures across the world can benefit from the efforts invested in certain parts of the project. Engineered open-source software applications could be incorporated into larger international community solutions. Adjacent to DiSSCo infrastructures globally, are also expected to make use of the information generated by the project in the short and midterm.

FAIR DATA

Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

All project outputs will be fully marked up through the use of a combination of generic document metadata and community-linked metadata schemata. This shall enable both the discoverability of the outputs across communities, as well as more targeted discoverability based on community-specific requirements. Each produced output shall receive a PID, based on the handle system (e.g. handle.net or DOI) and deposited in an internationally certified repository.

The draft set of metadata associated with each of the project's outputs are provided below. The metadata schema used will be further adjusted to improve the FAIR level of each of the outputs.

Table 3. User-entered fields to be used to annotate submitted project outputs, along with referenced namespace (where available) and short description. The metadata will improve the level of FAIR across all project outputs.

Field	Reference	Description
title	https://schema.datacite.org/meta/kernel-4.3/metadata.xsd	A name or title by which a resource is known.
authors	https://schema.datacite.org/meta/kernel-4.3/metadata.xsd	The main researchers involved working on the data, or the authors of the publication in priority order. May be a corporate/institutional or personal name.

contributors	https://schema.datacite.org/meta/kernel-4.3/metadata.xsd	The institution or person responsible for collecting, creating, or otherwise contributing to the development of the dataset.
publisher	https://schema.datacite.org/meta/kernel-4.3/metadata.xsd	The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource. This property will be used to formulate the citation, so consider the prominence of the role. In the case of datasets, "publish" is understood to mean making the data available to the community of researchers.
publicationYear	https://schema.datacite.org/meta/kernel-4.3/metadata.xsd	Year when the data is made publicly available. If an embargo period has been in effect, use the date when the embargo period ends.
embargo	https://schema.datacite.org/meta/kernel-4.3/metadata.xsd	If an embargo period has been in effect, use the date when the embargo period ends.
abstract	DiSSCo specific, but related to DataCite	abstract / short description for the resource
subject	https://schema.datacite.org/meta/kernel-4.3/metadata.xsd	Subject, keywords, classification codes, or key phrases describing the resource.
keywords	DiSSCo specific, but related to DataCite	„deliverable“, „milestone report“, „software pilot“ etc. plus scientific disciplines
language	https://schema.datacite.org/meta/kernel-4.3/metadata.xsd	Primary language of the resource. Allowed values are taken from IETF BCP 47, ISO 639-1 language codes.
rights	https://schema.datacite.org/meta/kernel-4.3/metadata.xsd	Any rights information for this resource. Provide a rights management statement for the resource or reference a service providing such information. Include embargo information if applicable. Use the complete title of a license and include version information if applicable.
license	Zenodo or DiSSCo specific	A reference to a well-known licence
resourceType	https://schema.datacite.org/meta/kernel-4.3/include/datacite-resourceType-v4.xsd	The type of a resource. You may enter an additional free text description.
format	https://schema.datacite.org/meta/kernel-4.3/metadata.xsd	Technical format of the resource. Use file extension or MIME type where possible.
modificationDate	DiSSCo specific, but related to DataCite	Date of the document's last modification
summaryOfModification	DiSSCo specific	Short description of the changes since creation or last modification

creationDate	DiSSCo specific, but related to DataCite	Date of the creation of the document's version
version	https://schema.datacite.org/meta/kernel-4.3/metadata.xsd	Version number of the resource. If the primary resource has changed the version number increases.
projectReference	DiSSCo specific[1]	Synthesys+, DiSSCo Prepare, ICEDIG
WP	DiSSCo specific	Work Package number of the respective project
task	DiSSCo specific	Task number of the respective project
deliverableNumber	DiSSCo specific	Number of official deliverable in the respective project
disseminationLevel	DiSSCo specific	Indicator for the target groups
documentCollection	DiSSCo specific	Tag of the collection the document belongs to

All project outputs will also be discoverable through the DiSSCo knowledge hub, to be developed during the project and subsequently maintained by the DiSSCo RI.

Products' naming convention will be further defined and shall provide enough semantic annotation to the outputs so that humans can quickly understand and provide the ability to reference the output to the relevant DiSSCo Prepare Task.

Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why clearly separating legal and contractual reasons from voluntary restrictions.

All data and outputs of DiSSCo Prepare will be made publicly available by default. The joint-owners of each one of the products will assign by default, and following the corresponding with the Consortium Agreement provisions, open-access licences to every project output. Open access licenses are limited to the use of outputs for any non-commercial activity.

The joint-beneficiaries can during the project agree to further open an output to include commercial use or transfer the intellectual ownership of the output.

Personal information collected throughout the project, under the provisions of GDPR, will remain confidential information and will not become publicly accessible. All relevant articles of the European GDP Regulation will also be applicable for this category of data collected or generated.

How will the data be made accessible (e.g. by deposition in a repository)?

All data will be deposited to an internationally certified public repository. Additional guidance will be issued to all beneficiaries, based on the recommendations of the DiSSCo Technical team and the beneficiaries of DiSSCo Prepare work package 5 during the early stages of the project. These recommendations shall provide step-by-step directions on how to deposit data generated/collected by each beneficiary.

[1] *DiSSCo specific metadata fields will be further detailed as part of the WP5 – Knowledge base development*

What methods or software tools are needed to access the data?
All project's outputs will be easily accessible and re-usable through standard desktop software applications. File formats expected to be generated are provided in table 2.

Is documentation about the software needed to access the data included?
The technical documentation shall be deposited along with the source code to open access git repositories. User documentation shall be deposited to generic repositories, following the project's practices for deposition of generated data/outputs to open access public repositories (e.g. Zenodo).

Is it possible to include the relevant software (e.g. in open source code)?
A dedicated DiSSCo git repository will host and provide access to source code developed throughout the project. Additional guidance shall be issued to all project's beneficiaries on the procedure to be followed for depositing and documenting software.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.
Data shall be deposited to an open access public and internationally certified repository. The DiSSCo knowledge hub shall provide an additional layer for discoverability of all the project's outputs, making use of the extended documents' metadata profiles (community-specific).

Have you explored appropriate arrangements with the identified repository?
The project management team shall further investigate the suitability of the Zenodo open public repository against other repositories and provide guidance to the beneficiaries on which solution is to be adopted for the deposition of all data and generated by the project material.

If there are restrictions on use, how will access be provided?
All joint-owners of the project's outputs shall provide an open access license. Restriction to non-commercial is already agreed through the project's Consortium Agreement. Creative Commons licenses will be preferred where applicable (documents, datasets etc), whilst a more detailed analysis on the software licensing scheme shall provide details on software licensing.

Is there a need for a data access committee?
The project's Executive Board in its role as a controller of key project outputs (According to the agreed upon Responsibility Assignment Matrix of the project), shall take up the role of a data access committee, where needed. Decisions of the Executive Board cannot contradict the relevant provisions in the Consortium Agreement nor any of the complementary provisions of this data management plan. When needed, consultation will be sought by the Technical Advisory Board and the Technical Team of DiSSCo.

Are there well described conditions for access (i.e. a machine readable license)?
Where possible (e.g. repository capabilities), machine readable and machine actionable licenses will be attached to the deposited project outputs.

How will the identity of the person accessing the data be ascertained?
The majority of the project outputs are to be published openly and identification of users will not be pursued. For sensitive information, such as personal data collected or generated by the project, closed access systems will be used. In particular the DiSSCo project management platform will ensure the use of Authorisation and Identification of all users. Access restrictions will be imposed to all personal information collected, based on the user agreement provisions made at the time of collections of such information (according to GDPR). Additional identification, when needed for the

purposes of better service provision, will be performed through an ORCID compatible infrastructure.

Making data interoperable

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Common data and metadata vocabularies, standards or methodologies will be used as available and appropriate to the content and types of digital object, including ISO 19115/19139 for geographic information; ISO 1806 for dates; ISO 3166 for country codes; Darwin Core (DwC), Access to Biological Collections Data (ABCD) and its extension for geosciences (ABCDEFG), and generic standards such as Exif and IIIF for images. Project outputs will be deposited to open access repositories using an augmented (see Table 3) DataCite metadata schema (https://schema.datacite.org/).

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

Formally ratified and community accepted Standards shall be used, wherever possible, to improve the level of FAIR across the project's outputs. Details of such standards have been given above. The DiSSCo knowledge hub will provide a full list of the standards used across the generated, by the project, datasets and other outputs.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Wherever possible, full mapping against known ontologies and vocabularies will be provided for all the fields used to describe project's outputs and datasets. When such ontologies do not exist, mapping will be done on a higher class.

Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

As mentioned above, all outputs will be issued an open access license for all but commercial re-use. This is pursuant to the provisions made in the project's Consortium Agreement. Joint owners shall evaluate the possibility of issuing licenses for commercial use on a case-by-case basis.

Software built during the project by project beneficiaries shall follow a licensing policy that encourages the project beneficiaries to use and build on software produced by DiSSCo and its contributing members. The recommended default choice is therefore a permissive license (e.g., either MIT or Apache V2.0). Further guidance on the licensing policy will be issued by the DiSSCo Technical Team and the DiSSCo Technical Advisory Board.

When will the data be made available for reuse? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

The DiSSCo Prepare project beneficiaries will open up all datasets and outputs generated as soon as possible and no embargo period is provisioned. In cases where such need arises by one of the joint-owners, the Executive Board will be asked to approve the request. Any embargo period imposed, however, shall not hinder the development of the DiSSCo Research Infrastructure.

Are the data produced and/or used in the project usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
All outputs of the project shall be made readily available for use by the DiSSCo Research Infrastructure as represented by a) the future legal entity of the infrastructure, b) the national nodes and c) any of the DiSSCo facilities.

Further to this, and given the expected usefulness of the outputs. Data shall persist and made available to any third party for reuse and according to the assigned licenses.

During the project the project council will further establish processes for the potential transfer of ownership of intellectual property generated throughout the project to the legal entity of DiSSCo to be established.

How long is it intended that the data remains re-usable?
Data shall remain accessible in the foreseeable future.

Are data quality assurance processes described?
Data quality assurance processes are embedded in the processes described by the project's responsibility assignment matrix. This matrix clearly describes all the involved parties and their role in controlling and advising the output owners (performers). The RAM is embedded in the managerial practices of the project and described in the Consortium Agreement.

ALLOCATION OF RESOURCES

What are the costs for making data FAIR in your project?
Costs associated with making all project outputs FAIR include a) Article Processing Charges for open access publications and b) Costs related to infrastructure maintenance of the DiSSCo knowledge hub.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).
APCs and infrastructure costs occurring during the project will be covered by a dedicated project budget. Costs that occur after the project ends, will be taken on by the DiSSCo RI core budget.

Who will be responsible for data management in your project?
Data management falls under the Management work package and is the overall responsibility of the project co-ordinator. The project Executive Board, the DiSSCo Technical Advisory Board and the Technical Teams will be also consulted, ad-hoc, on issues related to data management.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?
The DiSSCo RI will take on the responsibility for persistent access to the project outputs (including the DiSSCo knowledge hub and project website. The associated costs have already been included in the multiannual work and budget plan of the infrastructure, beyond the timeframe of the project.

DATA SECURITY

Security provisions will be implemented at the institutional level. Where applicable, Data Protection Officers (in institutions) shall ensure the protection of sensitive (including personal) information collected. In case where such information is deposited in Consortium platforms, such as the project management platform of the project, the project management team will ensure that restricted access is available. The project management platform follows strict authorisation and authentication practices.

ETHICAL ASPECTS

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

All surveys circulated by the project shall provide specific information on the extent and exact purpose of use of the collected data, according to the provisions of the European GDP Regulation.

The project beneficiaries shall ensure that data collected through online platforms, will be held at European servers, based on the assurances provided by the platform operators.