ELIXIR-CONVERGE: Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services (871075)

Laura Portell Silva, Salvador Capella-Guttierez, Pinar Alper, Teresa d'Altri, Adam Hospital, Espen Aberg, Daniel Faria, Anne-Françoise Adam-Blondon

Deliverable D5.2 Report on the first two DMP processes

Project Title (Grant agreement no.):	ELIXIR-CONVERGE: Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services (871075)		
Project Acronym (EC Call):	ELIXIR-CONVERGE (H2020-INFRADEV-2018-2020)		
WP No & Title:	WP5 Demonstrator Projects		
WP leader(s):	Anne-Françoise Adam-Blondon, Alfonso Valencia, Salvador Capella-Gutierrez		
Deliverable Lead Beneficiary:	20 - CNR		
Contractual delivery date:	31/07/2021	**Actual delivery date:**	30/07/2021
Delayed:	No		
Partner(s) contributing to this deliverable:	BSC, SIB, HITS, INRAE, CNRS, MTA, CNR, UNILU, UIB, INESC-ID, IGC, UU, UL, CRG		
Authors: Laura Portell (BSC), Salvador Capella-Gutierrez (BSC), Pinar Alper (UNILU), Teresa d'Altri (EGA), Adam Hospital (IRB Barcelona), Espen Åberg (UiB/UiT), Daniel Faria (INESC-ID), Anne-Françoise Adam-Blondon (INRAE) **Contributors:** Olivier Collin (CNRS), Nils P Willassen (UiB/UiT), Paulette Lieby (CNRS), Anastasis Oulas (CING), Evangelos Pafilis (HCMR), Niclas Jareborg (NBIS), Nazeefa Fatima (UiB/UiO), Wolmar Nyberg Åkerström (UU/SU), Ernesto Picardi (CNR) **Acknowledgments (not grant participants):** Cyril Pommier (EMPHASIS, ELIXIR Plant Science community, INRAE); Sebastian Beier (ELIXIR Plant science community, IPK)			
Reviewers:	ELIXIR-CONVERGE Management Board (MB) members.		

Log of changes

DATE	Mvm	Who	Description
31/05/2021	0v1	Salvador Capella-Gutierrez (BSC) Laura Portell (BSC)	Initial version
21/07/2021	0v2	Salvador Capella-Gutierrez (BSC) Laura Portell (BSC)	Sent to PMU after incorporating internal WP feedback
22/07/2021	0v3	Nikki Coutts (ELIXIR Hub)	Circulated to the MB for final review before submission
30/07/2021	1v0	Nikki Coutts (ELIXIR Hub)	Final version to be uploaded into EC Portal

Table of contents

1. Executive Summary

The goal of ELIXIR Converge WP5 is to assess the capacity of ELIXIR and its national nodes to assist users' projects in implementing Data Management Plans (DMPs) in their projects at a EU scale. To do

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

that, six demonstrator use-cases were created and they are described in the previous deliverable (D5.1)[1].

This deliverable includes a description of the main resources that can be used to assess researchers in the process of creating DMPs for at least two use-cases including a gap analysis. The DMPs are related to the demonstrator use-case #1 (Plant sciences) and #3 (Marine metagenomics). Moreover, specific aspects for use-cases #4 (Human data) and #6 (Biomolecular simulation) are also added.

The main activities described in this deliverable are related to the ELIXIR Research Data Management Kit (RDMkit)[2] and the Data Stewardship Wizard (DSW)[3] in connection with ELIXIR Converge WP3 activities. For all use-cases except use-case 5 (Toxicology data), a 'Your domain' page in the RDMkit has been assembled with the information necessary to consider when managing research data in their corresponding domains. Also, a 'Tool assembly' page in the RDMkit for the plant sciences, marine metagenomics and human data use-cases have been added. Finally, the DSW has been used to perform the initial description of the first DMPs including a gap analysis.

Since ELIXIR Converge started, the work of WP5 has been centered in understanding the needs of the demonstrator use-cases to develop, if necessary, customized DMPs for each of them considering the domains they are part of. Also, once this characterization was done, several resources have been created to help users in the development of their DMPs. Indeed, these demonstrators are expected to serve as examples for researchers developing their own DMPs.

The domain and tool assemblies pages added in the RDMkit are ready to use internally in ELIXIR after the beta-release in February, 2021. They provide a valuable resource for researchers and data stewards in the life sciences to help them in their data management processes. The Knowledge Models (KMs), the central component of the DSW, are under periodic revision to incorporate outcomes from ELIXIR Converge including specific aspects identified by the use-cases in order to complete their content and release an official version of them.

The final objective is to provide refined DMP versions for the remaining 4 use-cases. This will allow us to cover a broad range of data-driven demonstrators, with its associated challenges, and assess how the general process scales up when applied to other communities served by ELIXIR. All these efforts will contribute to harden the work developed in other ELIXIR Converge WPs like the RDMkit and the DSW (WP3), the dedicated training materials and capacity building actions (WP2) as well as to provide reference information to the data management experts network (WP1) and external members to the project, e.g. Industry, in WP4.

[1] 10.5281/zenodo.4674490
[2] https://rdmkit.elixir-europe.org/
[3] https://ds-wizard.org/

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

2. Contribution toward project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives/key results:

Objective no. / Key Result no. Description	Contributed to:
Objective 1: Develop a sustainable and scalable operating model for transnational life-science data management support by leveraging national capabilities **(WP1, WP5**)	
Key Result 1.1: Established European expert network of data stewards that connect national data centres and similar infrastructures and drive the development of interoperable solutions following international best practice, including national interpretations of the General Data Protection Regulation (GDPR)	**Yes**
Key Result 1. 2: Development of joint guidelines and common toolkit that are adopted into funder recommendations, with support available nationally and in local languages	**Yes**
Key Result 1.3: The catalogue of successful national business models incorporated into national strategies	**No**
Key Result 1.4: The developed "sustainable and scalable operating model for transnational life-science data management support" is adopted into national ELIXIR Node	**Yes**
Objective 2: Strengthen Europe's data management capacity through a comprehensive training programme delivered throughout the European Research Area **(WP2, WP6)**	
Key Result 2.1: A comprehensive ELIXIR Training and Capacity building programme in Data Management, directed at both data managers and ELIXIR users, and connected to the national training programmes in Data Management in the ELIXIR Nodes and prospective ELIXIR Member countries.	**No**
Key Result 2.2: Development of a collective group of trainers that support scalable deployment of Data Management training across ELIXIR Nodes.	**No**
Key Result 2.3: A substantial cohort of data managers, Node coordinators and researchers with specific data management skills, business planning and knowledge of transnational operations across the ELIXIR Nodes	**No**
Objective 3: Align national data management standards and services through a sustainable, scalable and cost-effective data management toolkit **(WP2, WP3, WP5**)	
Key Result 3.1: Assemble a full-stack harmonised common toolkit comprising all aspects of data management: from data capture, annotation, and sharing; to	**Yes**

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

integration with analysis platforms and making the data publicly available according to international standards.	
Key Result 3.2: Provide exemplar toolkit configurations for prioritised demonstrators to serve as templates for future use.	**Yes**
Key Result 3.3: Establish national capacity in using as well as updating, extending and sustaining the toolkit across the ERA.	**No**
Key Result 3.4: Enable 'FAIR at source' practice for data generation, and analytical process pipeline implementation by flexible deployment of the toolkit in national operations	**No**
Objective 4: Align national investments to drive local impact and global influence of ELIXIR (**WP4,WP6**)	
Key Result 4.1: Development of a Node Impact Assessment Toolkit based on RI-PATHS methodology.	**No**
Key Result 4.2: Adoption of Impact assessment in ELIXIR Nodes, supported by Node coordinators network and feedback on applicability from dialogues with national funders.	**No**
Key Result 4.3: Creation of national public-private partnerships and industry outreach where open life-science data and services stimulate local bioeconomy	**No**
Key Result 4.4: Growth in reach, impact and engagement of stakeholder communication assessed by established ELIXIR Communications metrics	**No**
Key Result 4.5: Initiating and advancing discussions on Membership (EU and international) or strategic partnerships (international countries) following ELIXIR-CONVERGE workshops.	**No**

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

3. Introduction

Data Management Plans (DMPs) are key for good research data management, as they describe how research data is handled before, during and after the end of the project.[4] The overall goal of ELIXIR Converge WP5 is to assess the capacity of ELIXIR and its national nodes to assist users' projects in implementing DMPs in their projects at a EU scale. A set of six very diverse demonstrator use-cases were selected and are addressed for this purpose:

1. Harmonised FAIR plant genotype & phenotype data management toolkit for Europe
2. Reproducible, comparable and FAIR Epitranscriptomics
3. Common Data management plans for the marine metagenomics Community
4. Federated access to human genomics data: GDPR
5. FAIR encoding and access to Toxicology data
6. FAIR organisation of biomolecular simulation information

In previous steps, in task 5.1, the different demonstrators use-cases were analysed in light of the process necessary to implement their DMPs and to propose a categorization based on users' needs. During that effort, the needs in terms of DMP of the demonstrator use-cases were described, and a project categorization was presented (Deliverable D5.1[5]).

3.1 Scope of Deliverable

This deliverable includes a description of the main resources that can be used to assess researchers in the process of creating DMPs for at least two of the use-cases in close collaboration with ELIXIR Converge WP3. In addition, This work is completed by a gap analysis. The DMPs are related to the demonstrator use-case #1 (Harmonised FAIR plant genotype & phenotype data management toolkit for Europe) and #3 (Common Data management plans for the marine metagenomics Community). Also, considering their broad reach, specific contributions for the demonstrator use-cases #4 (Federated access to human genomics data: GDPR) and #6 (FAIR organisation of biomolecular simulation information) are also included in this deliverable.

The main activities described in this deliverable are related to the ELIXIR Research Data Management Kit (RDMkit)[6] and the Data Stewardship Wizard (DSW)[7]. The RDMkit is a website-based toolkit designed to guide life sciences scientists and data stewards in their efforts to better manage their research data while the DSW has been designed to assist their users to efficiently compose data management plans (DMPs) for their research projects.

[4] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
[5] 10.5281/zenodo.4674490
[6] https://rdmkit.elixir-europe.org/
[7] https://ds-wizard.org/

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

- For all use-cases except use-case 5 (Toxicology data), a 'Your domain' page in the RDMkit has been assembled with the information necessary to consider when doing data management in their corresponding domains.
- A 'Tool assembly' page in the RDMkit for the marine metagenomics and human data use-cases have been also added.
- The DSW has been used to perform the initial description of the first DMPs including a gap analysis.

These activities provide feedback to WP3 contributing to the development of processes to enrich and extend the RDMkit (Task 3.2).

3.2 Relationship with other WPs

- ELIXIR Converge WP1 Data Management experts contributed in the best practices and guidelines in the 'Your domain' pages in the RDMkit for some of the demonstrator use-cases added.
- ELIXIR Converge WP2 members contributed in the creation of dedicated training materials and capacity building actions.
- ELIXIR Converge WP3 RDMkit members contributed to the assembly of the 'Your domain' pages in the RDMkit for some of the demonstrator use-cases, as well as the marine metagenomics 'Tool assembly' page.
- ELIXIR Converge WP4 members contributed in providing reference information to the external members to the project, e.g. Industry.
- ELIXIR Converge WP7 Federated EGA members helped in the design and gap analysis of the human data use-case.

3.3 Methodology

The methodology of work for this deliverable was developed with the representatives of the demonstrator use-cases mainly during the monthly meetings and side exchanges with some relevant ELIXIR communities (e.g. Plant Science), experts of DSW and ELIXIR Converge WP3 partners.

The tool assembly pages related to the demonstrators use-cases and the work on the DSW were developed through two complementary approaches:

1. The tool assembly contentathon organized by WP3 (see D3.2[8]) building on the detailed description of the demonstrators with the methodology described in D5.1[9] and completed afterwards.

[8] https://zenodo.org/record/5139897#.YQAyoRNKj0o
[9] https://zenodo.org/record/4674491#.YQPBBRNKj0o

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

2. A working session with DSW experts on the adaptation of the DSW features in the context of the demonstrators use-cases.

4. Description of work accomplished

The different activities done for this deliverable are focused on creating resources for researchers of the different domains comprising the demonstrator use-cases. This effort will provide life sciences researchers and data stewards with reference DMPs, which can assist them in the process of creating their own DMPs. This includes the pages in the RDMkit and also adaptation of the general Knowledge Model (KM) of the DSW. Since most of the tasks described for this deliverable were done in close collaboration with ELIXIR Converge WP3, the description of work accomplished has shared many references with the work described in D3.2[10].

4.1 RDMkit 'Your domain' pages

A page in the RDMkit was created for all the demonstrator use-cases except for Toxicology data, which is under construction. These pages are under 'Your domain' category on the RDMkit and they describe the needs and considerations that need to be considered to implement good data management practices depending on the domain which they belong to.

In the RDMkit 'Your Domain' pages, challenges specific to domains such as data types, species or areas are highlighted together with the main considerations and solutions that can be used to overcome them. The scope of the 'Your Domain' content varies depending on the context. For some domains it can be very large (e.g. Human Data) and for others it is narrower (e.g. Epitranscriptomics).

4.2 RDMkit 'Tool assembly' pages

In addition to the 'Your Domain' pages, two tool assemblies were also added in the RDMkit corresponding to three of the demonstrator use-cases: Plant Sciences, Marine Metagenomics and Human data. Tool Assemblies are examples of combining tools to provide data management across the stages of the research data management lifecycle or a subset of stages, and the distillation of assembly patterns (blueprints) to guide how tools can be combined.

In order to write these pages, members of WP5 participated in the contanthons organized by WP3 to extend the tool assemblies pages in the RDMkit. These contentathons happened in May and June 2021 and they had participants from WP1, WP2, WP3 and WP5.

[10] https://zenodo.org/record/5139897#.YQAyoRNKj0o

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

4.3 Data Stewardship Wizard Knowledge Models

For the generation of DMPs for this deliverable, the DSW was used as a tool that brings together data stewards and researchers to efficiently compose DMPs for their research projects. The DSW is built around a hierarchical KM that is used to lead data stewards through a decision tree to help them choose the right tools, resources and practices when making DMPs. Since the information relevant for the DMPs might vary between domains, these KMs can be customized for each of the use-cases.

To create these customized KMs, a DSW instance for the ELIXIR Converge project was created (URL). Then, a specific session with DSW experts and representatives of the different use-cases was held, where the general KM for DMPs was analysed to detect areas that can be improved for each of the use-cases. During this session, four of the demonstrator use-cases were analysed (Plant Sciences, Marine Metagenomics, Human Data and Biomolecular Simulation) and a method was set up to enrich the DSW KMs.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

5. Results

5.1 RDMkit 'Your domain' pages

5.1.1 Use-case #1: Plant sciences

In this page[11], challenges specific to the plant sciences domain are highlighted together with the main considerations and solutions that can be used to overcome them. In the plant sciences, the main challenge addressed is data integration that needs a standardized way to identify the heterogeneous ensemble used in various experiments comprising plant varieties, their accessions obtained from a genebank or a laboratory and all the samples derived from them. The description of the plant material at infra-specific levels (e.g. subspecies, variety, accession name) often lacks consensus. Recently, guidelines have been developed to solve this problem by the plant sciences community.

Another particularity of this domain is the absence of central deposition databases for certain important data types, in particular data coming from plant phenotyping experiments. Whereas datasets from plant omics experiments are typically deposited in global deposition databases for that type of experiment, those from phenotyping experiments remain in institutional or at best national repositories. Finding, accessing and interconnecting plant phenotyping data is therefore based on utilizing a common metadata standard and on following the FAIR principles to facilitate the emergence of federations of interoperable information systems.

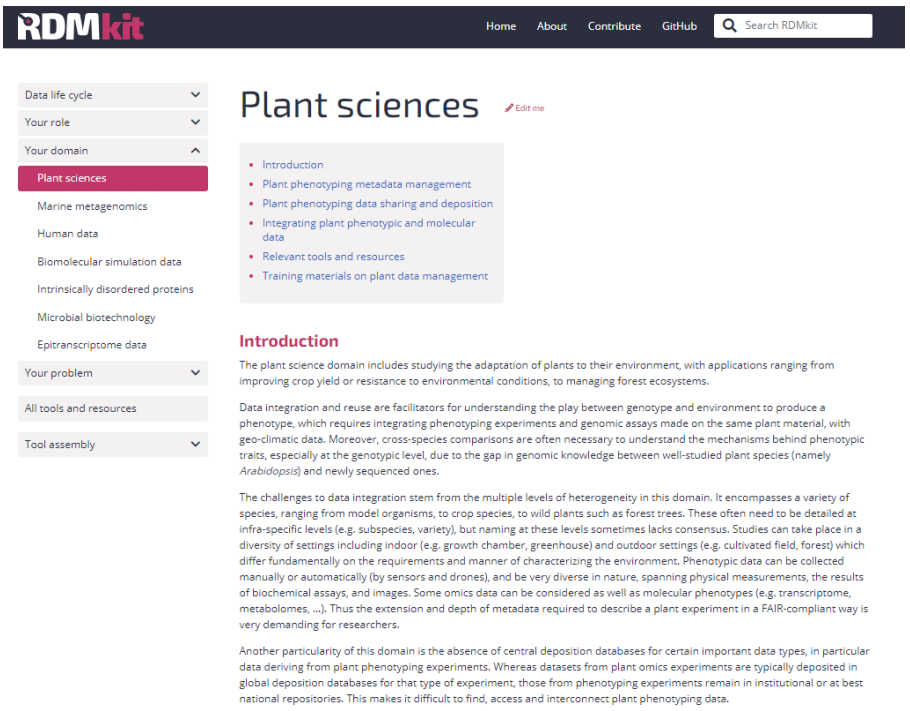


Figure 1. Plant science domain page in the RDMkit

[11] https://rdmkit.elixir-europe.org/plant_sciences.html

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

5.1.2 Use-case #2: Epitranscriptome data

Epitranscriptome modifications are appearing as important factors to fine tune gene expression and regulation in a variety of organisms and experimental conditions. Thus the detection of RNA modifications requires bioinformatics pipelines and the use of computing intensive tools to handle the huge amount of available deep transcriptome sequencing experiments.

A fruitful organization of data and computational workflows is an important aspect for the data management of this domain. It has been decided to tackle those challenges by FAIRing those components as well as providing recommendations for this particular domain. In this page[12], challenges that are characteristic to the epitranscriptome data domain are mentioned together with the main considerations and solutions that can be used as a solution for them.

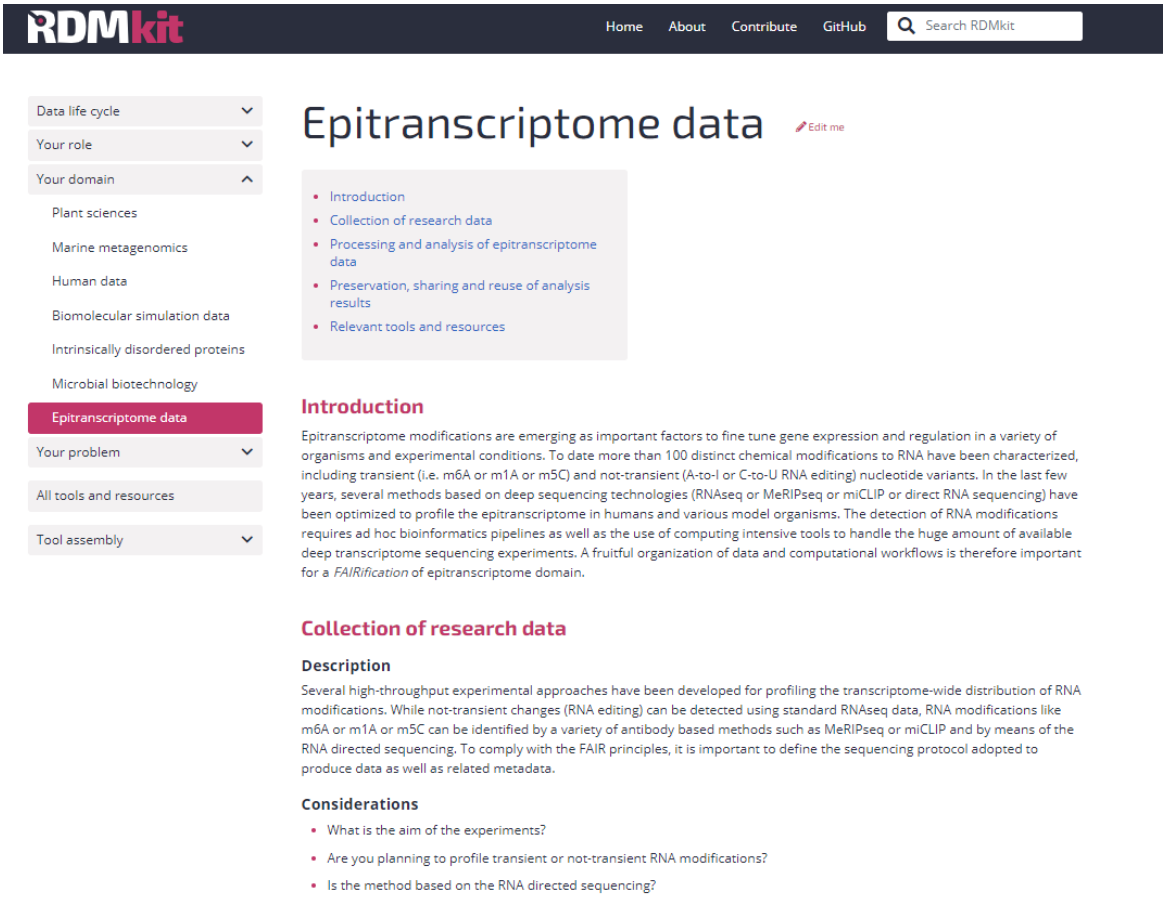


Figure 2. Epitranscriptome data domain page in the RDMkit

[12] https://rdmkit.elixir-europe.org/epitranscriptome_data.html

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

4.1.3 Use-case #3: Marine metagenomics

The marine metagenomics domain[13] has very specific challenges when considering research data management. They have been listed in this entry for the RDMkit as well as different considerations and potential solutions for those challenges.

The marine metagenomics domain characterizes for using large datasets that require access to substantial storage and High-Performance Computing (HPC) for running complex and memory-intensive analysis pipelines, and therefore are difficult to handle for typical end-users and are beyond the resources available to many service providers. To enable research under those conditions, the community has put the focus on the use and re-use of metagenomics data. To be able to share and reuse metagenomics datasets, they have to be in compliance with the FAIR principles which relies entirely on recording rich metadata in all the data lifecycle steps from sampling to data analysis.

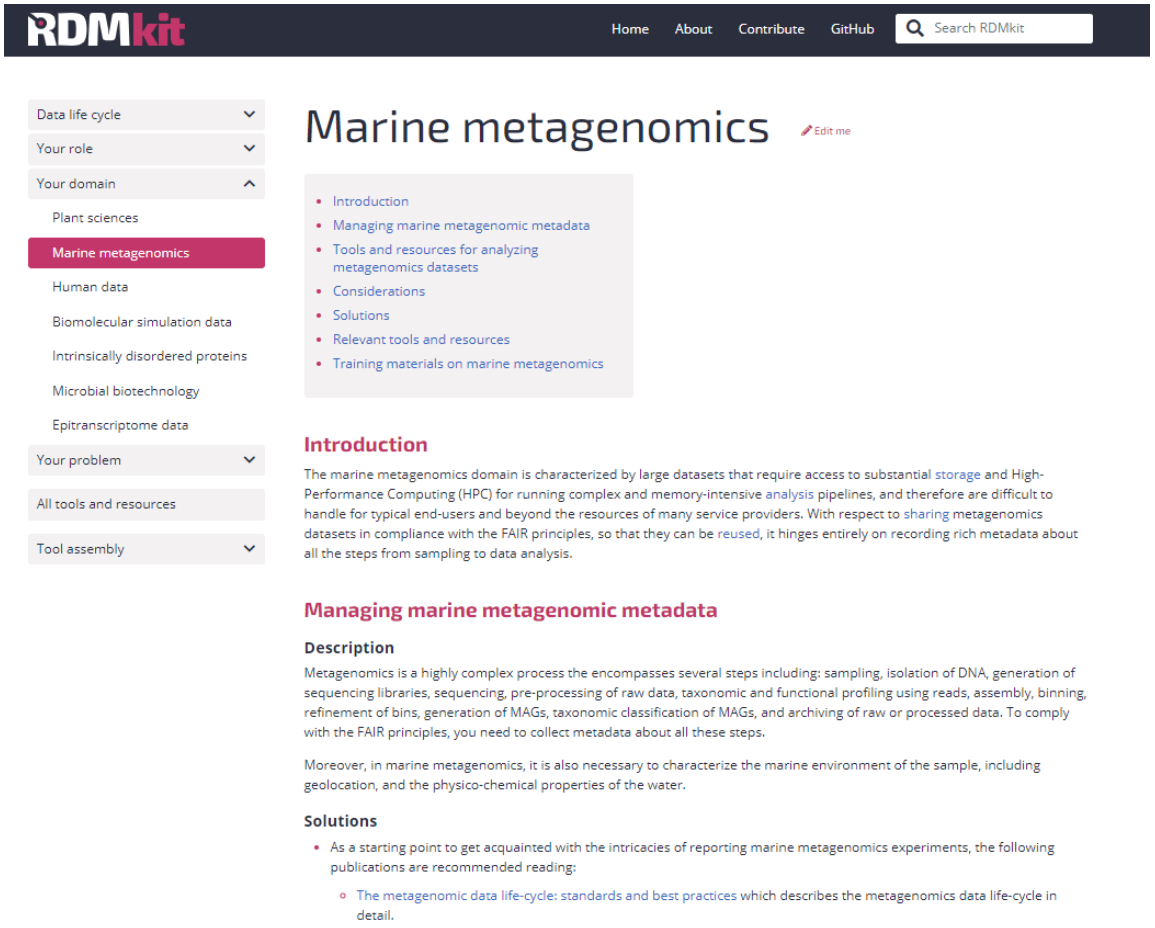


Figure 3. Marine metagenomics domain page in the RDMkit

[13] https://rdmkit.elixir-europe.org/marine_metagenomics.html

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

4.1.4 Use-case #4: Human data

There are additional challenges that need to be considered when doing research with data derived from human individuals. Those particular challenges associated with the fact of working with sensitive data are described in this RMDkit page. Much of the topics discussed on this page refer to the EU General Data Protection Regulation (GDPR) as it is a central piece of legislation that affects basically all research done on human subjects in the EU and on individuals residing in the EU. Also, much of the information on this page[14] is of a general nature when it comes to working with human data but an additional focus is on human genomic data and the sharing of such information for research purposes.

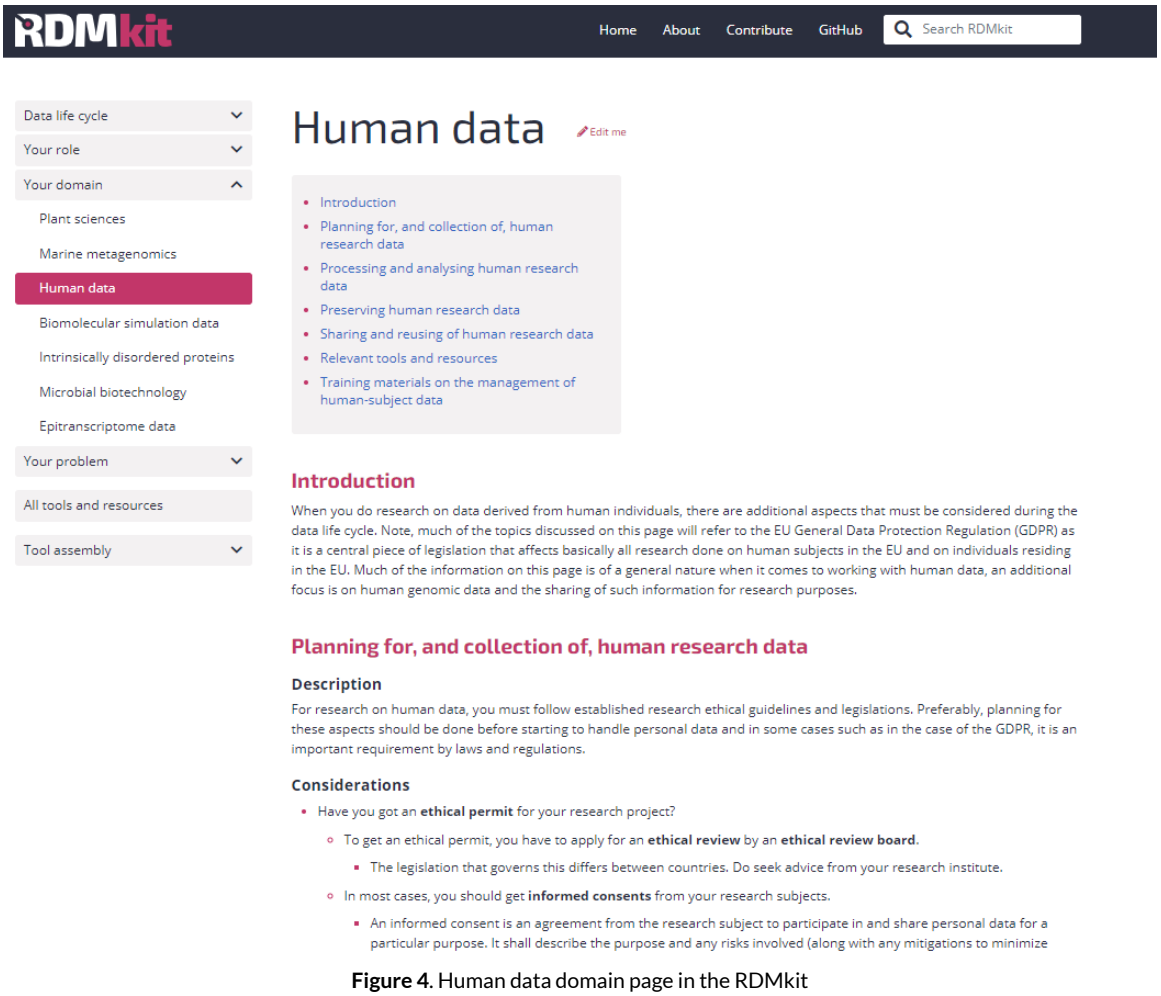


Figure 4. Human data domain page in the RDMkit

[14] https://rdmkit.elixir-europe.org/human_data.html

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

4.1.5 Use-case #6: Biomolecular simulation data

Biomolecular simulation methods are demonstrating rapidly growing impact in areas as diverse as biocatalysis, drug delivery, biomaterials, biotechnology, and drug or protein design. Yet the field only relatively recently started to store and share (bio)simulation data to be reused for new, unexpected projects, and started discussions about their biomolecular simulation data being more FAIR. In this page[15], challenges specific to the biomolecular simulation domain are highlighted together with several current possibilities moving towards the direction of FAIRifying this type of data in the broad sense.

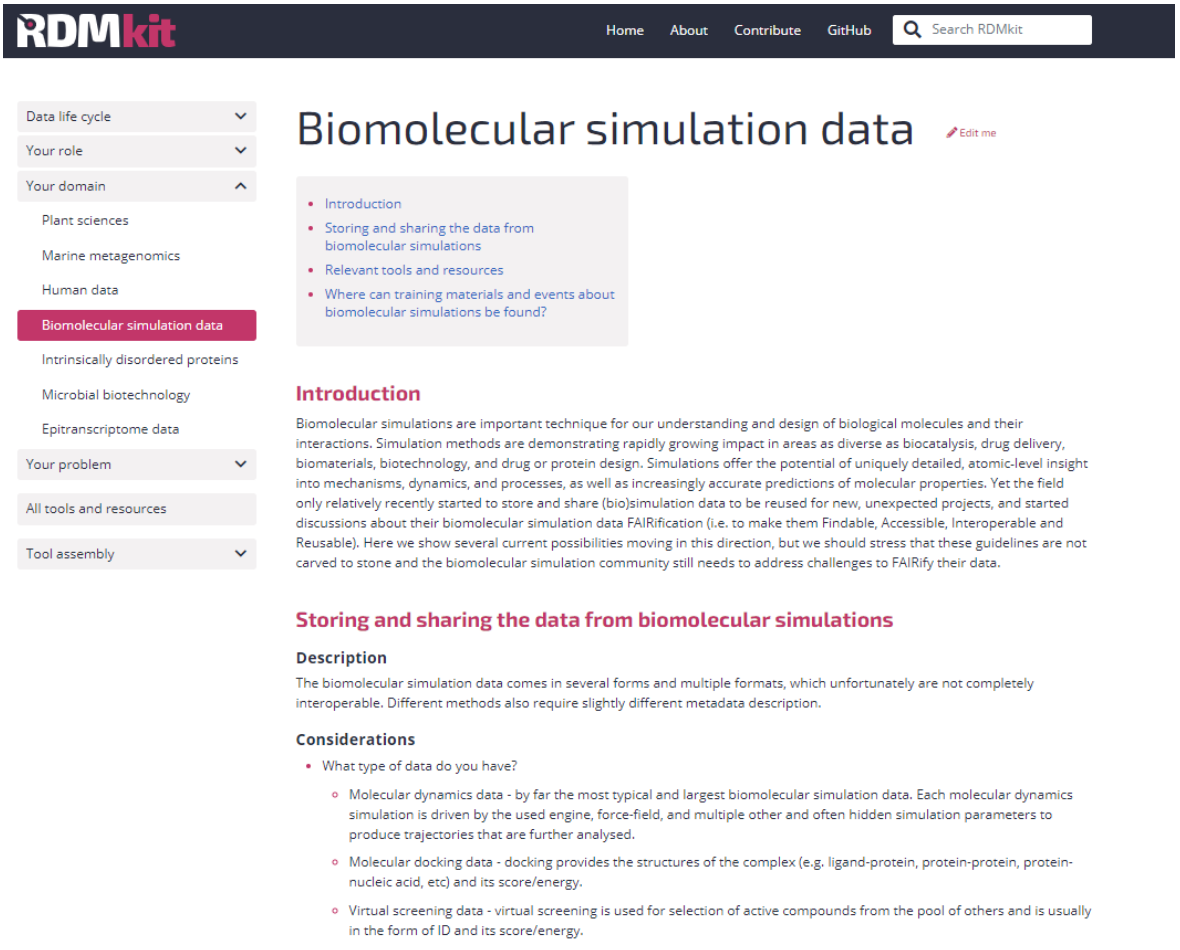


Figure 5. Biomolecular simulation page in the RDMkit

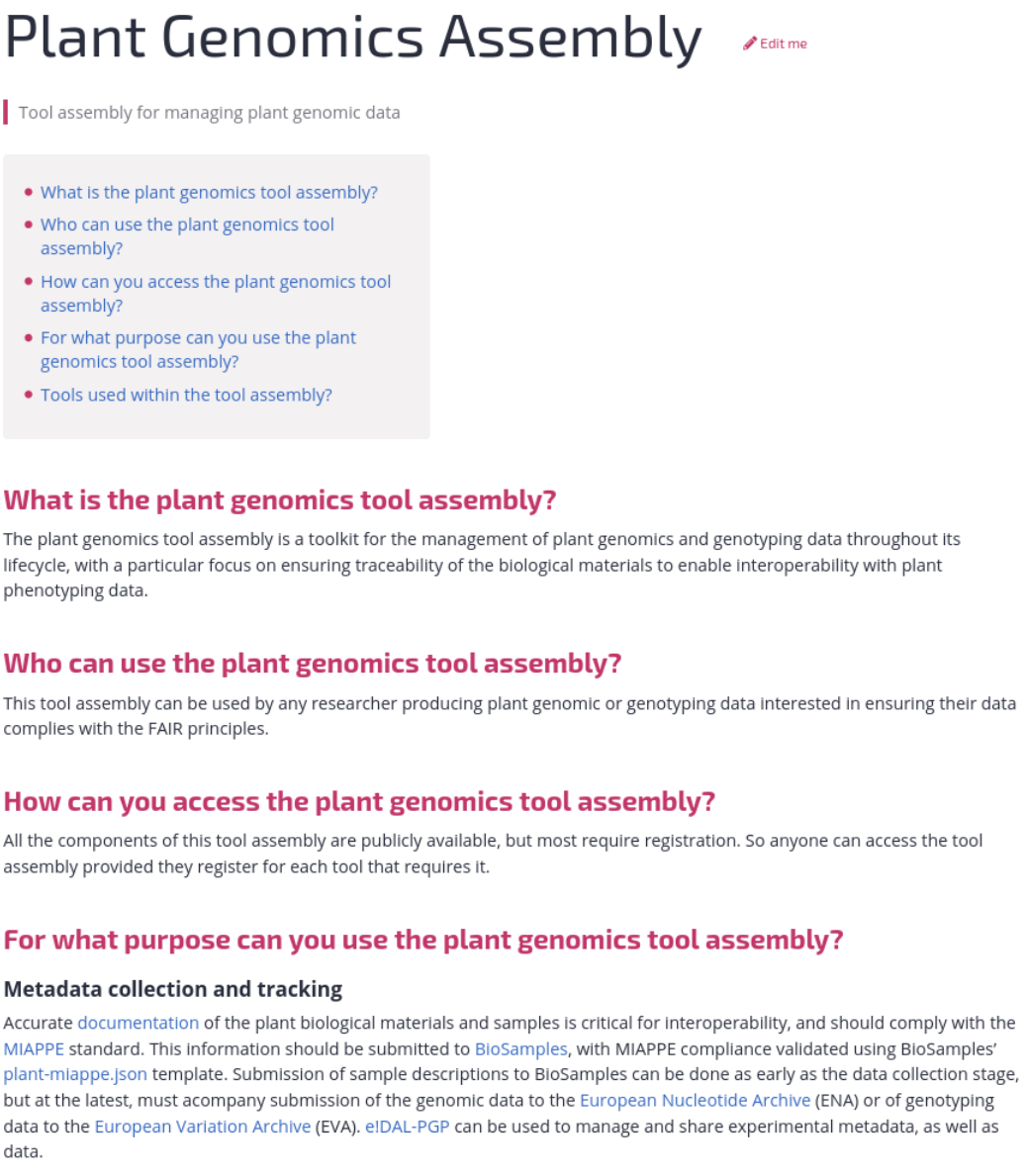
[15] https://rdmkit.elixir-europe.org/biomolecular_simulation_data.html

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

5.2 RDMkit 'Tool assembly' pages

5.2.1 Use-case #1: Plant sciences tool assembly

This page[16] is created to help users from the plant sciences community in the management of genomic data during the different steps of the data life-cycle. It has particular focus on ensuring traceability of the biological materials to enable interoperability with plant phenotyping data.



Plant Genomics Assembly

Edit me

Tool assembly for managing plant genomic data

- What is the plant genomics tool assembly?
- Who can use the plant genomics tool assembly?
- How can you access the plant genomics tool assembly?
- For what purpose can you use the plant genomics tool assembly?
- Tools used within the tool assembly?

What is the plant genomics tool assembly?

The plant genomics tool assembly is a toolkit for the management of plant genomics and genotyping data throughout its lifecycle, with a particular focus on ensuring traceability of the biological materials to enable interoperability with plant phenotyping data.

Who can use the plant genomics tool assembly?

This tool assembly can be used by any researcher producing plant genomic or genotyping data interested in ensuring their data complies with the FAIR principles.

How can you access the plant genomics tool assembly?

All the components of this tool assembly are publicly available, but most require registration. So anyone can access the tool assembly provided they register for each tool that requires it.

For what purpose can you use the plant genomics tool assembly?

Metadata collection and tracking

Accurate documentation of the plant biological materials and samples is critical for interoperability, and should comply with the MIAPPE standard. This information should be submitted to BioSamples, with MIAPPE compliance validated using BioSamples' plant-miappe.json template. Submission of sample descriptions to BioSamples can be done as early as the data collection stage, but at the latest, must acompany submission of the genomic data to the European Nucleotide Archive (ENA) or of genotyping data to the European Variation Archive (EVA). eIDAL-PGP can be used to manage and share experimental metadata, as well as data.

Figure 6. Plant genomics tool assembly page in the RDMkit

[16] https://rdmkit.elixir-europe.org/plant_genomics_assembly.html

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

5.2.2 Use-case #3: Marine metagenomics tool assembly

This tool assembly[17] aims to provide an example of combining tools for management of marine genomic research data throughout a project's data life cycle. In this case, the tool assembly is established at national level, especifically to the Norwegian node of ELIXIR.

The toolkit, developed by students and researchers in Norway, contains resources and software tools for both data management (Planning, Processing, Storing and Sharing), data analysis and training. It is built on the Norwegian e-Infrastructure for Life Sciences (NeLS) tool assembly of ELIXIR Norway and the Marine Metagenomics Platform (MMP).

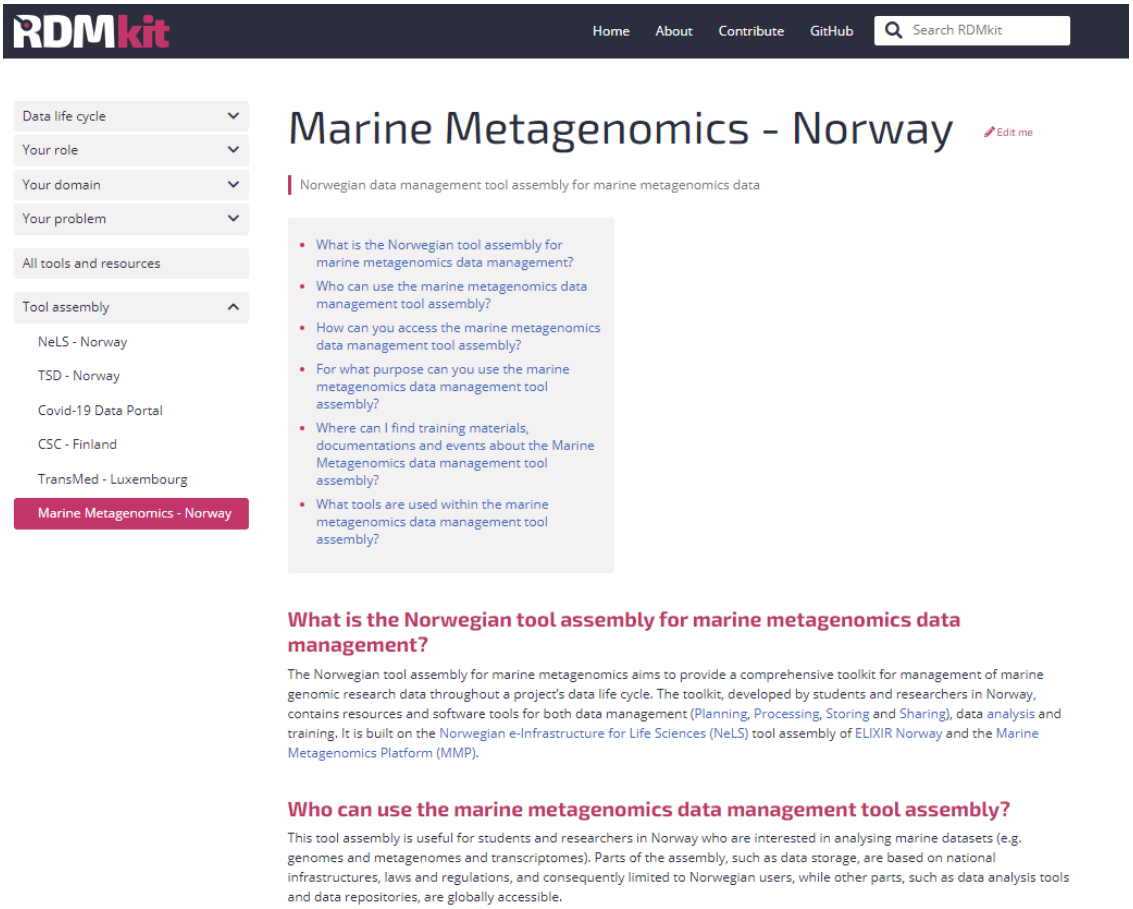


Figure 7. Norwegian marine metagenomics tool assembly page in the RDMkit

[17] https://rdmkit.elixir-europe.org/marine_metagenomics_assembly.html

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

5.2.3 Use-case #4: TransMed tool assembly for working with sensitive human data

The TransMed tool assembly[18] is also a blueprint showing how a collection of tools can be combined to support data life cycle management in clinical and translational projects. This is especially relevant for those projects handling sensitive data associated with phenotypic and genotypic patients.

The TransMed data and computing tool assembly is an infrastructure provided by ELIXIR Luxembourg for clinical and translational projects. TransMed assembly provides the tools for managing ongoing projects that often require the management of cohort recruitment, and processing of samples, data and metadata. This entails GDPR-compliant and secure data collection, storage, curation, standardisation integration and analysis of clinical data and associated molecular, imaging and sensor/mobile data and metadata.

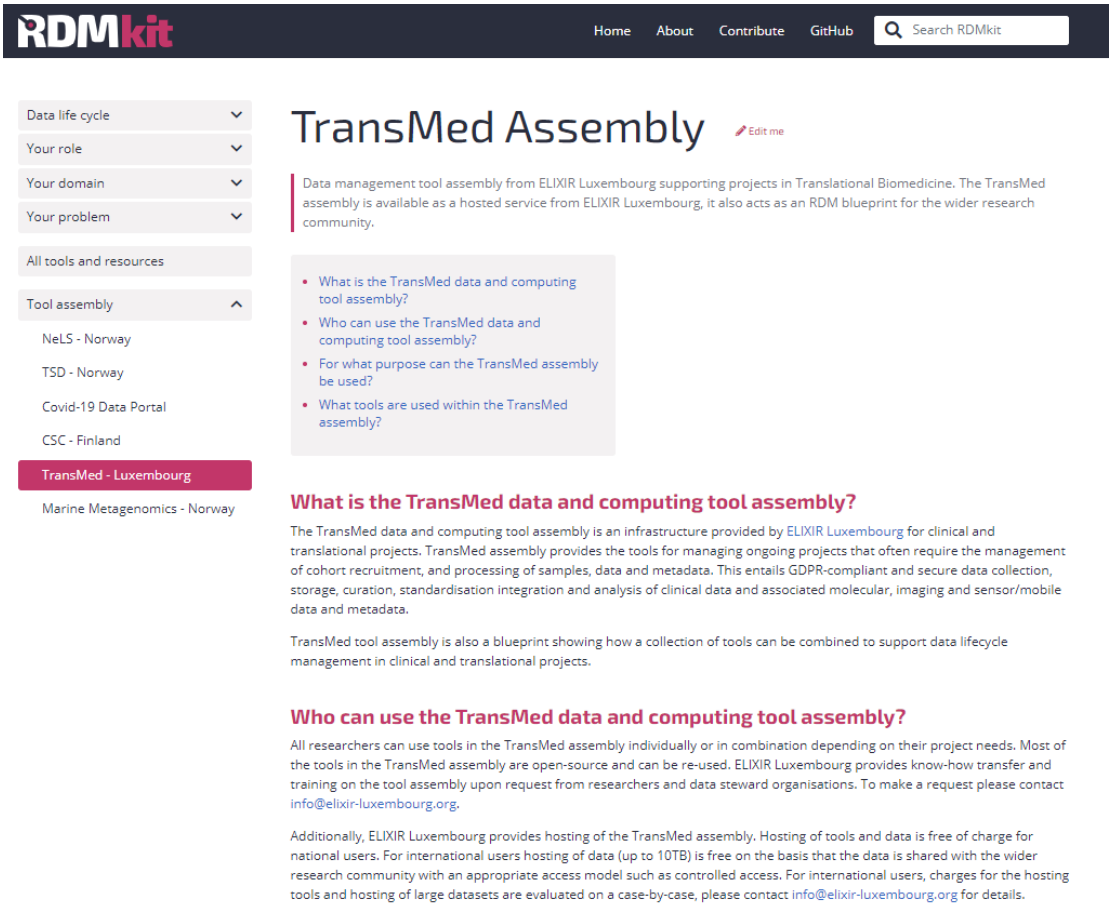


Figure 8. Luxemburg TransMed tool assembly page in the RDMkit

[18] https://rdmkit.elixir-europe.org/transmed_assembly.html

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

5.3 Data Stewardship Wizard Knowledge Models

The KMs in DSW contain the knowledge about what should be asked and how to get the necessary information from users to generate a DMP. The KMs follow a tree-like structure and are organized into Chapters, Questions, Answers and additional resources. Therefore, the KMs are the cornerstone to generate the set of questions that a given user is present to.

In the Converge DSW instance, there are two KMs available for consideration (see figure 8). One of them is a more general one called "*Common DSW Knowledge Model*" and another specific for life sciences called "*Life Sciences DSW Knowledge Model*".

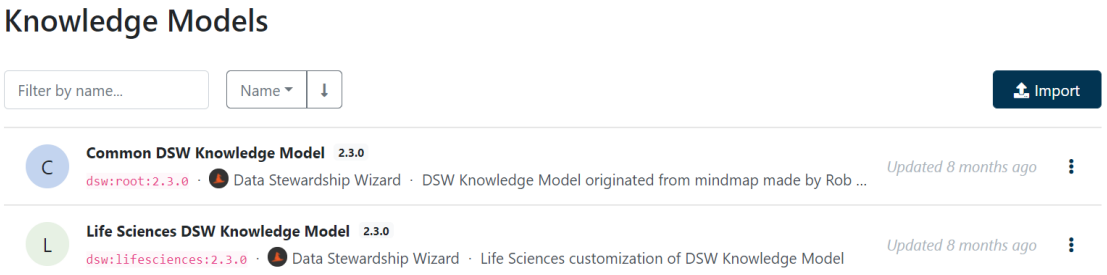


Figure 9. Data Stewardship Wizard Knowledge Models.

For this deliverable, the Common DSW KM was used as a template to be evaluated, commented and refined by each of the use-cases. To initiate this work, a gap analysis was done to evaluate the missing aspects of the Common DSW KM in order to make it more specific for each of the use-cases' domains. In figure 8, it is possible to see the projects that were created in the DSW for the documentation of this gap analysis. For the marine metagenomics use-case, the norwegain instance of the DSW was used and therefore it is not shown in the picture.

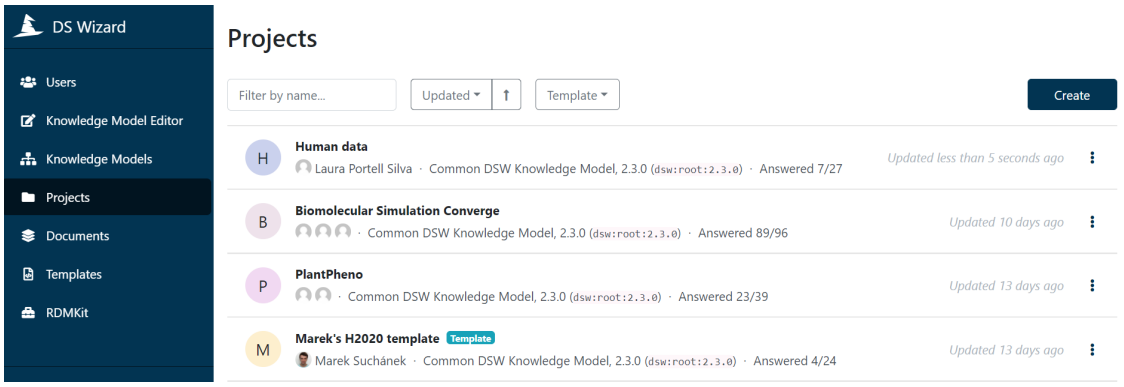


Figure 10. Data Stewardship Wizard projects

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

5.3.1 Use-case #1: Plant sciences

For the Plant science demonstrator use-case, the focus was on the most specific type of data: the data derived from plant phenotyping experiments.

- **Reusing data:** The first step is to gather existing passport data and identifiers associated with the plant material. This information can be obtained from genebanks if the material was obtained from a gene bank via their own catalog or via international catalogs of plant genetic resources (e.g. EURISCO for Europe and Genesys at the global level); advanced research plant material is also often generated by laboratories with internal identifiers and metadata sets. For the moment, the catalogs of genetic resources are poorly connected to the scientific community, not referenced in FAIRsharing and therefore not available in the Common DSW KM.

- **Creating and collecting data:** the (meta)data created will correspond to data about the seed lots or samples generated for phenotyping experiments, data about the protocols of measurements and the phenotypic data themselves.

  It is often difficult to understand at what level the questionnaire is and it would be useful to have a section summarizing the different data sets that will be described in the preserving and archiving sections and associated data and metadata.

  The plant specific metadata standard MIAPPE, although in FAIRsharing it was not in the proposed list.

  The section on the provenance of the data is very oriented towards human health applications and could be adapted in a template dedicated to plants.

- **Processing data:** this step was not the focus of the plant use-case.

- **Interpreting data:** the section on interoperability with other data is very relevant to the plant use-case but might have to be checked with researchers for the use cases of integrations that are proposed (are they the most frequents? Is the phrasing understandable?)

- **Preserving data:** the datasets are described here and the difference in terms of information with the "collecting data" section is to be clarified to avoid repetitions (e.g. in terms of ontologies). Again a section somewhere describing briefly the datasets and the metadata and data associated could help.

As a conclusion, the Common DSW KM could very well be used to generate a specific template dedicated to the plant science community by filtering out or re-phrasing some items, which are

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

oriented towards human health data, into plant science specific questions as well as by adding some plant specific resources in the lists, e.g. MIAPPE.

This exercise also points out some improvements needed in the plant science community, in particular in terms of specific reference repositories around the plant material and the phenotyping data. The conclusion is clear, the community needs better cataloging strategies and start using such catalogues connected with their respective data repositories.

5.3.2 Use-case #3: Marine metagenomics

For the marine metagenomics use-case, the DMP model is based on the *Life Sciences DSW Knowledge Model - ELIXIR Norway localization* which is a customization of the *Life Sciences DSW Knowledge Model*. This KM is hosted in the Norwegian DSW instance but is publicly available from the Marine Metagenomics RDM tools assembly page. At present, this KM is established at national level to be compliant with several European DMP models, such as Science Europe and Horizon 2020. It also encompasses information required at the national level such as the NeLS storage request form. The model covers all the needs of norwegian users and partly that of other european users at the moment. An effort to develop a more general European version is ongoing with the objective to understand the process of dedifferentiation from a specific KM towards a general one. This result can provide feedback on the process of creating specific versions and/or branches of the Common KM.

The most important gaps to achieve this dedifferentiation mainly lie in the solutions for data storage, sharing and computation, which are built in connection with national services and infrastructure. Such services and infrastructures are delivered by ELIXIR Norway and are described in the Norwegian e-Infrastructure for Life Sciences (NeLS) tool assembly. Suggestions in the KM describing these solutions would not always be relevant or even accessible to non-norwegian users and will have to be reworked for pan-european users.

Of note users can select different tags to create their DMPs covering questions relevant for European projects, e.g. H2020 and HE. The KM also makes ubiquitous suggestions on existing workflows, databases, vocabularies, ontologies and file formats specifically directed towards marine metagenomics. These can be isolated into a marine metagenomics project DMP by selecting the Marine MMP tag.

In addition the KM includes a short section on data sensitivity and general ethical guidelines even though this aspect, in most cases, will be less important for marine metagenomics projects than for the human related ones.

The practical aspects around the implementation of a Marine Metagenomics - European KM are currently being explored. The main challenge is that there is no easy way to migrate the ELIXIR Norway location DSW back to the parent model. The selection of databases and tools specific to

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

marine metagenomics will be expanded to include additional important resources such as MGnify and the Marine Microbial Gene Catalogue (MarCat) among others.

5.3.3 Use-case #4: Human data

To perform the gap analysis for the human data use-case, it was evaluated that the considerations mentioned in the RDMkit Human Data page[19] were reflected in DSW. The main gaps that were found in the DSW KM for this particular use-case are:

- There are no questions on Ethical Considerations, and Legal Considerations appear to come up serendipitously in "data processing" and onwards. Therefore, separate sections are needed, one that covers the ethical and another that covers the legal aspects that the project should consider during the data management planning phase, like general data protection aspects and their implementation through the Data Protection Impact Assessment (DPIA)[20].

- The DSW is trying to direct the user to making the data open by applying every method possible (aggregation or anonymisation). We think that controlled-access and registered-access should be brought forward and presented as valid legal options to share human-subject data.

- Under the "Giving access to data" section, the DSW touches upon a number of use restrictions that could be applicable to data. For the sake of completeness, it might be useful to point the user to the Global Alliance for Genomics and Health (GA4GH) Data Usage Ontology (DUO) codes and ask the researchers if there will be any further restrictions and/or data access scenarios not being contemplated in such ontology.

- GDPR and anonymisation/pseudonymisation is found under "data processing" within an information security related section "Is the risk of information loss, leaks and vandalism acceptably low?" We believe this is a bit late to introduce these concepts. Perhaps in the "Ethical and Legal Considerations" section or under "Creating and collecting data" researchers should be asked whether researcher will pseudonymise the data at source.

- In the question "Is the risk of information loss, leaks and vandalism acceptably low?", would be good to add a close by section, on whether national platforms for secure processing are known e.g. Ga4GH compatible clouds, GA4GH data security toolkit, ISO/IEC 27001.

When completing the questions of the KM, it makes the user understand that to have open data is always the best option. In the case of human data, most of the time it doesn't have to be open, it can be controlled access it wouldn't be a problem.

[19] https://rdmkit.elixir-europe.org/human_data.html
[20] https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/when-data-protection-impact-assessment-dpia-required_en

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

5.3.4 Use-case #6: Biomolecular simulation data

- **Metadata**: A choice between Dublin Core Metadata, DataCite Metadata or DDI Metadata is given. In the case of the Biomolecular simulation field, most of the time it is not needed to use any metadata standards and therefore people from this field might not know any of them.

- **Checksums / Swaps**: might be too specific to ask if checksums or check file swaps of the data are saved.

- **Standard Operating Procedure (for file naming)**: giving such a complicated name to ask if standard nomenclature is used, makes it more difficult to understand what the actual question means.

- **Provenance**: It is possible to choose between lab notebooks, electronic lab notebooks, and "other". In computing, most of the time it will be "other". In this case, the documentation, or appropriate reference are missing to know what other provenance standards could be used.

- **Saving data**: Options are Object Store, Relational database (DB), Graph DB, and Triple Store. Other options including noSQL and non-relational DBs are missing.

- **Identifiers**: In this case, options are Handle, DOI, ARK, URL and other. It is possible that identifiers for the entries in a dataset are not used. Entries in biomolecular simulation can be N files of a single molecular dynamics simulation, including path, topology, analysis, etc. In some projects an URL with a persistent identifier is included, but not for all of them.

- In general, a certain sense of repetition especially in regards to data types / formats when you get to step 5.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

6. Conclusions

Since ELIXIR Converge started, the work of WP5 has been centered in understanding the needs of the demonstrator use-cases to develop customized DMPs for each of them considering the domains they are part of. Also, once this characterization was done, several resources have been created to help users in the development of their DMPs. This includes the "Your domain" and "Tool assembly" pages in the RDMkit and an evaluation towards personalized knowledge models for the DSW.

The exercise to build tool assemblies for the demonstrators use-cases was facilitated by the work achieved to deliver domain pages and adding questions about the national implementations of tool assemblies. Therefore, some of the resources needed along the data life-cycle, such as storage and computing resources, are provided at the institutional or country level. The Marine Metagenomic use case is an example on that aspect: a Norway-grounded tool assembly was delivered (NeLS) and the partners responsible of the demonstrator use-case are working on approaches to give a more generic tool assembly.

The DSW is overall a well received system and is very well suited to generate DMPs for all demonstrator use-cases but some important sections would be necessary (ethical and legal aspects) or helpful (overview of the data sets and data and metadata associated) to be added. To solve this, a modular approach could be considered. This can imply an additional chapter in the Common KM with pre-requisites, e.g. have you covered all ethical and legal considerations for the type of data you will managed along the project life cycle, or links to external complementary resources, e.g. DAISY for the ethical and legal issues management for human sensitive data. Some pre-listed resources or formats are missing in the DSW KM and could be added, together with cross links between the domain pages of the RDMkit and DSW.

The domain and tool assemblies pages added in the RDMkit are ready to use internally in ELIXIR after the beta-release in February, 2021. They provide a valuable resource for researchers and data stewards in the life sciences to help them in their data management processes. The KMs are under periodic revision to incorporate outcomes from ELIXIR Converge including specific aspects identified by the use-cases in order to complete their content and release an official version of them.

After a year and a half of work, the ELIXIR Converge WP5 has built a productive team that includes data management experts as well as representatives of all the 6 demonstrator use-cases. The division of the tasks among members and the participants on internal activities has been very successful. Even with the broad difference between use-cases, WP5 managed to develop common methods to work and produce great and valuable work in close collaborations with other WPs in the project.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

7. Impact

The activities associated with this deliverable have had a direct impact across different components of the projects, e.g. the RDMkit and the DSW. In the case of RDMkit, pages for the "Your domain" section have been added for all the demonstrator use-cases except one, which gives a lot of value to the toolkit. From now on, users from the domains of plant sciences, marine metagenomics, human data, biomolecular simulations and epitranscriptomics will be able to address their specific challenges when doing data management by just having a look at these pages. Also, the tools listed in the "Your domain" pages as solutions, have been flagged in bio.tools with the domain name. This categorization will allow researchers and data stewards that work in these particular domains to be able to find solutions to their data management problems in an easier and faster way.

Globally, the work achieved so far by WP5 has also given a push towards completing common catalogs with important resources in the context of each demonstrator use-cases. This effort has to be pursued, with a global vision, e.g. making sure that important resources not maintained by ELIXIR members are visible, to give more impact to the RDMkit.

In addition, the tools assemblies for plant sciences, marine metagenomics and human data give a very good insight on how national nodes use a combination of their tools for different steps of their data life-cycle, which can help other nodes to do the same.

8. Next Steps

The perspective of the work is:

1. Finish the RDMkit domain page for the Toxicology data use-cases.
2. Add a tool assembly page in the RDMkit for the rest of the use-cases.
3. Generate a knowledge model adjusted to each of the use-cases, and evaluate the best strategy to release it. Current strategies are two: incorporate use-case driven particularities as part of the branches of the common model or create adjusted KM per domain.
4. Continue the collaboration with the ELIXIR Converge WP2 on the definition of the training and capacity building activities.

9. Deviation from Description of Action

The Description of Action proposal included a description of the first two DMP processes. However, more work regarding the rest of the use-cases was done and therefore stated in this deliverable. All demonstrator use-cases except one have a page in the RDMkit and three of them a tool assembly. Also, a gap analysis of the KM in the DSW has been done for four of the use-cases already. Once these KMs are completed, they will serve as a template for the users to generate DMPs.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.