



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

Anne-Françoise Adam-Blondon, Pinar Alper, Salvador Capella-Gutierrez, Daniel Faria, Laura Portell Silva, Adam Hospital, Nils-Peder Willassen

### **► To cite this version:**

Anne-Françoise Adam-Blondon, Pinar Alper, Salvador Capella-Gutierrez, Daniel Faria, Laura Portell Silva, et al.. ELIXIR-CONVERGE: Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services (871075): Deliverable D5.1 Categorisation of the pilot projects. [Research Report] INRAE. 2021. hal-03310082

**HAL Id: hal-03310082**

**<https://hal.inrae.fr/hal-03310082>**

Submitted on 30 Jul 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC0 - Public Domain Dedication 4.0 International License

## Deliverable D5.1 Categorisation of the pilot projects

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		
Deliverable Lead Beneficiary:	12 - INRAE		
Contractual delivery date:	31/01/2021	Actual delivery date:	06/04/2021
Delayed:	Yes		
Partner(s) contributing to this deliverable:	VIB, UCY, UOCHB, HITS, DTU, UT, BSC, MTA, UCD, WEIZMANN, CNR, UNILU, UIB/UIT, UU, UNIMAN, UCAM		
<b>Authors:</b> A-F Adam-Blondon (INRAE), P Alper (UNILU), S Capella-Gutierrez (BSC), D Faria (INESC-ID), L Portell Silva (BSC), A Hospital (IRB), N-P Willassen (UiT),; <b>Contributors:</b> Frederik Coppens (VIB); Danny Ben-Avraham (WEIZMANN); Olivier Collin (CNRS); Niclas Jareborg (NBIS); Anastasis Oulas (CING); Evangelos Pafilis (HCMR); Gregoire Rossier (SIB); Venkata Satagopam (UNILU); Celia Van Gelder (DTU); Ulrike Wittig (HITS) <b>Acknowledgments (not grant participants):</b>			
Reviewers:	ELIXIR-CONVERGE Management Board (MB) members.		

### Log of changes

DATE	Mvm	Who	Description
30/11/2020	0v1	Anne-Françoise Adam-Blondon (INRA)	Initial version
19/2/2021	0v2	Anne-Françoise Adam-Blondon (INRA)	Sent to PMU after incorporating internal WP feedback
19/2/2021	0v3	Nikki Coutts (ELIXIR Hub)	Circulated to the MB for final review before submission

4/3/2021	0v4	Anne-Françoise Adam-Blondon (INRA)	MB comments addressed
8/4/2021	1v0	Nikki Coutts (ELIXIR Hub)	Final version to be uploaded into EC Portal

## Table of contents

<b>1. Executive Summary</b>	<b>2</b>
<b>2. Contribution toward project objectives</b>	<b>3</b>
<b>3. Introduction</b>	<b>5</b>
<b>4. Description of work accomplished</b>	<b>7</b>
<b>4.1 Development of methods and templates to collect more information about the demonstrator use-cases</b>	<b>7</b>
4.1.1 Users needs in relation with the demonstrator use-cases	7
4.1.2 Description of the resources and expertises in relation with the demonstrator use-cases	8
<b>4.2 Method for collecting the data</b>	<b>9</b>
<b>5. Results</b>	<b>9</b>
<b>5.1 The personas and their requirements</b>	<b>9</b>
5.1.1. The data consumers	9
5.1.2 The data producers	10
5.1.3 The workflow productors	10
5.1.4 Data owners/managers/controllers	11
5.1.5 Data curators	11
5.1.6 Conclusions	11
<b>5.2 Expertise, resources and training: ELIXIR's support and gaps</b>	<b>12</b>
<b>6. Conclusions</b>	<b>13</b>
<b>7. Impact</b>	<b>14</b>
<b>8. Next Steps</b>	<b>15</b>
<b>9. Deviation from Description of Action</b>	<b>15</b>

## 1. Executive Summary

Task 5.1 aims to analyse the different demonstrators in light of the process necessary to implement their DMPs and to propose a categorization based on users' needs. This effort will provide input to WP1 with respect to needs in terms of experts and resources across ELIXIR Nodes (Task 1.3, Business plans). It is expected that some of them will be generic whereas others will be very specific to projects, which is an important factor for establishing a strategy towards the sustainability of such a service in all its diversity.

D5.1 therefore describes the needs in terms of DMP of the demonstrator use-cases and proposes a project categorization from the point of view of:

- Area of expertise
- The resources they use / toolkit
- Training needs

The first challenge was to develop a method to describe the needs of the users and their current practices in terms of data management as a prerequisite to the development of an adequate toolkit. The method adopted with the guidance of WP2 and WP3 relied on the description of users stories, and scenarios for different types of personas. In addition, a table was also developed to collect knowledge on the current practices and existing resources in terms of data management with guidance from WP1 and WP3. The six demonstrator use-cases are divided into two waves to facilitate having a broad diversity of data types, potential users and variable expertise and, at the same time, have first mature demonstrator use-cases that can start providing input to other WPs in the projects. The Plant and the Marine Metagenomics demonstrator use-cases were used as testers for the method developed as they are the two selected ones for the first wave. Federated access to human genomics data belongs to the second wave of demonstrator use-cases. However, considering its broad reach, it has been partially developed for this deliverable.

User stories and scenarios with personas were developed by WP5 partners on the Plant, Marine metagenomics and Federated access to human genomics data: GDPR use-cases. The table on existing tools and resources was also filled in by WP5 partners and target interviews of researchers for some of the demonstrator use-cases. Additional input was sought from the ELIXIR communities for the demonstrator use-cases mentioned before.

This method implemented to describe the demonstrator use-cases in detail has been efficient to isolate the specific context of each case, and to identify specific tools/resources in relation with the use-cases that can be used for data management.



The main aspect used to classify a new use-case is the type of biological material linked to the data that links to many specific requirements, e.g. legal and scientific metadata. The second is the relation of the personas with the data, e.g. consumers/producers, and the type of data, including workflows, models, and the various types of experimental data.

The landscape analysis of the resources already available in relation with the demonstrator use-cases and the different stages of a data management plan, showed that the guidelines developed by WP1 are very useful to clarify the scope of these stages before filling such a survey.

The personas/users related to the use-cases are very diverse but always include the researchers in various positions, e.g. data consumer, data producer, data owner, PI writing a proposal, and the data curators. This will be further developed in collaboration with WP2. Finally, the work achieved to deliver D5.1 facilitated the development of pages dedicated to the “domains” in the RDMKit website (<https://rdmkit.elixir-europe.org/>) for demonstrator use-case 1, 3, 4 and 6. In these pages, specific challenges linked with the demonstrator use-cases are highlighted, and dedicated resources to solve them are listed. These tools have also been flagged in bio.tools with the domain name.

## 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:
<b>Objective 1:</b> Develop a sustainable and scalable operating model for transnational life-science data management support by leveraging national capabilities ( <b>WP1, WP5</b> )	
<b>Key Result 1.1:</b> 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)	<b>Yes</b>
<b>Key Result 1.2:</b> Development of joint guidelines and common toolkit that are adopted into funder recommendations, with support available nationally and in	<b>Yes</b>



local languages	
<b>Key Result 1.3:</b> The catalogue of successful national business models incorporated into national strategies	No
<b>Key Result 1.4:</b> The developed “sustainable and scalable operating model for transnational life-science data management support” is adopted into national ELIXIR Node	No
<b>Objective 2:</b> Strengthen Europe’s data management capacity through a comprehensive training programme delivered throughout the European Research Area (WP2, WP6)	
<b>Key Result 2.1:</b> 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
<b>Key Result 2.2:</b> Development of a collective group of trainers that support scalable deployment of Data Management training across ELIXIR Nodes.	No
<b>Key Result 2.3:</b> 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
<b>Objective 3:</b> Align national data management standards and services through a sustainable, scalable and cost-effective data management toolkit (WP2, WP3, WP5)	
<b>Key Result 3.1:</b> Assemble a full-stack harmonised common toolkit comprising all aspects of data management: from data capture, annotation, and sharing; to integration with analysis platforms and making the data publicly available according to international standards.	Yes
<b>Key Result 3.2:</b> Provide exemplar toolkit configurations for prioritised demonstrators to serve as templates for future use.	Yes
<b>Key Result 3.3:</b> Establish national capacity in using as well as updating, extending and sustaining the toolkit across the ERA.	No
<b>Key Result 3.4:</b> Enable ‘FAIR at source’ practice for data generation, and analytical process pipeline implementation by flexible deployment of the toolkit in national operations	No
<b>Objective 4:</b> Align national investments to drive local impact and global influence of ELIXIR (WP4,WP6)	
<b>Key Result 4.1:</b> Development of a Node Impact Assessment Toolkit based on	No

RI-PATHS methodology.	
<b>Key Result 4.2:</b> Adoption of Impact assessment in ELIXIR Nodes, supported by Node coordinators network and feedback on applicability from dialogues with national funders.	<b>No</b>
<b>Key Result 4.3:</b> Creation of national public-private partnerships and industry outreach where open life-science data and services stimulate local bioeconomy	<b>No</b>
<b>Key Result 4.4:</b> Growth in reach, impact and engagement of stakeholder communication assessed by established ELIXIR Communications metrics	<b>No</b>
<b>Key Result 4.5:</b> Initiating and advancing discussions on Membership (EU and international) or strategic partnerships (international countries) following ELIXIR-CONVERGE workshops.	<b>No</b>

### 3. Introduction

The overall goal of WP5 is to assess the capacity of ELIXIR and its national nodes to assist users' projects in implementing data management plans (DMPs) in their concrete projects at a EU scale. A set of six very diverse demonstrator use-cases will be addressed for this purpose.

As a first step, Task 5.1 aims to analyse the different demonstrators in light of the process necessary to implement their DMPs and to propose a categorization based on users' needs. This effort will provide input to WP1 with respect to needs in terms of experts and resources across ELIXIR Nodes (Task 1.3, Business plans). It is expected that some of them will be generic whereas others will be very specific to projects, which is an important factor for establishing a strategy towards the sustainability of such a service in all its diversity.

D5.1 therefore describes the needs in terms of DMP of the demonstrator use-cases and proposes a project categorization from that point of view:

- Area of expertise
- The resources they use/toolkit
- Training needs

In addition to the initial plans, we decided to make a first mapping of resources (expertise, tools, training) that are already identified across ELIXIR's nodes and possibly beyond to support these needs in the light of their specialization to certain areas of research or, on the contrary, their genericity. This will improve the interactions with WP3 (development of the toolkit) and WP2 (development of training).

Six demonstrator use-cases are considered in WP5 in two waves and are described below.

**1. Harmonised FAIR plant genotype & phenotype data management toolkit for Europe** (1st wave; FR, DE, IT, PT, UK).

- Data types & Challenge: phenotyping data (tabular text, ISA-tab), genotyping data (SNP genotyping matrices in VCF either from genotyping arrays or GBS experiments, SSR type identification markers), includes data with restricted access.
- Key Deposition databases, standards and Interoperability resources: ENA, EVA, BioSamples, ISA-toolkit, MIAPPE, BrAPI, CropOntology.

**2. Reproducible, comparable and FAIR Epitranscriptomics** (2nd wave; IE, IL, IT, NO).

- Data types & Challenge: Available high-throughput sequencing data sets, e.g. deep RNA-seq of 25k samples; deep RNA-seq (MeRIP, miCLIP, ...) and Ribo-seq of 2.3k samples.
- Key Deposition databases, standards and Interoperability resources: Harmonised/benchmarked workflows, MIAME, BioSamples.

**3. Common Data management plans for the marine metagenomics Community** (1st wave; FR, DE, HU, IT, NO, PT, UK).

- Data types & Challenge: Large Metagenomics datasets including whole metagenomes shotgun data, metagenome assembled genome data, metatranscriptome, metabarcoding data
- Key Deposition databases, standards and Interoperability resources: Drive implementation of community standards across ELIXIR Nodes<sup>1</sup>

**4. Federated access to human genomics data: GDPR** (2nd wave; FI, FR, GR, IT, LU, NO, ES, SE, CH).

- Data types & Challenge: A typical ELIXIR Community member would be a research institute/medical centre that is willing to share their research data, but doing so under more restrictive conditions than general research reuse. Reusability of sensitive data in particular requires special effort from data users as well as data providers in the context of GDPR and its country-specific implementation in context of the 1M Genomes Declaration.
- Key Deposition databases, standards and Interoperability resources: Local EGA functionalities, Beacon, ELIXIR Beacon Network, Data Access Management tools, GDPR tools, ELIXIR AAI, GA4GH standards.

**5. FAIR encoding and access to Toxicology data** (2nd wave; DK, EMBL-EBI, ES).

- Data types & Challenge: Chemical structures, pharmacological data, toxicological data, data from clinical trials, information on drug side-effects, pharmacovigilance data, etc.

<sup>1</sup> <https://doi.org/10.1093/gigascience/gix047>



- Key Deposition databases, standards and Interoperability resources: ChEMBL, OLS, Transnational access to complex data, terminology mappings (OxO).

#### 6. FAIR organisation of biomolecular simulation information (2nd wave; ES, SE).

- Data types & Challenge: long coordinate files containing the evolution of 3D positions of atoms of the system across time; output of a wide variety of analysis tools applied to raw data, metadata describing system setup and the simulation parameters. Simulation data is typically stored locally in an undocumented manner, without any external curation and lacking associated metadata required for their reusability.
- Key Deposition databases, standards and Interoperability resources: Some initial attempts in the generation of usable ontologies (UMM, BigNASim).

## 4. Description of work accomplished

### 4.1 Development of methods and templates to collect more information about the demonstrator use-cases

The first challenge was to develop a method to describe the needs of the users on their current practices in terms of data management as a prerequisite to the development of an adequate toolkit. The six demonstrator use-cases are divided into two waves to facilitate having a broad diversity of data types, potential users and variable expertise and at the same time have first mature demonstrator use-cases that can start providing input to other WPs in the projects. In that sense, the Plant and the Marine Metagenomics use-cases were used as testers for the method developed as they are the two selected ones for the first wave. Federated access to human genomics data belongs to the second wave. However, considering its broad reach, it has been partially developed for this deliverable.

#### 4.1.1 Users needs in relation with the demonstrator use-cases

WP5 first developed user stories and scenarios involving different types of personas covering each demonstrator use-case. This method was proposed by Carole Goble (WP3) and Munazah Andrabi (WP3) who provided some documents of reference and definitions (see box 1). User stories were collected using a template developed by the GO FAIR initiative: Copy of Stocktaking GO FAIR Discovery IN - Use cases, infrastructure - for publication<sup>2</sup>

<sup>2</sup><https://docs.google.com/spreadsheets/d/1qK8sNP-zL9wVOWMrsX8dv2Tv5oia4sKk4pjLKBM9KMA/edit#gid=0>

**Box 1**

**Stakeholders** identifies those organisations and people who contribute to, impact or are impacted by CONVERGE toolkit training and experts.

**Persona** are the characters who will appear in our user stories. They represent our users and makers of the toolkits, the data steward experts and trainers and others. Many (but now all) will be drawn from the Stakeholders.

**User Scenarios** are fictitious “Alice and Bob” stories of accomplishing an action or goal via the Toolkit. The characters in the scenario are drawn from the Persona.

**User Stories** are high level requirements that could be translated later into the technical requirements for the toolkit and follow a strict template.

see also: [https://en.wikipedia.org/wiki/Persona\\_\(user\\_experience\)](https://en.wikipedia.org/wiki/Persona_(user_experience))

User stories per personas were developed for three demonstrator use-cases: (see: ELIXIR-CONVERGE User Stories\_v1<sup>3</sup>)

- Demonstrator use-case 1 - Harmonised FAIR plant genotype & phenotype data management
- Demonstrator use-case 3 - Common Data management plans for the marine metagenomics Community
- Demonstrator use-case 4 - Federated access to human genomics data: GDPR

To complete the description, scenarios involving personas were also described: : see ELIXIR-CONVERGE\_Demonstrator\_Scenarios\_v1<sup>4</sup>.

The description of the scenarios was easier and more natural than the table proposed by GO FAIR. In addition, it allows to describe the context in which the personas interact, absent in the GO FAIR template.

#### **4.1.2 Description of the resources and expertises in relation with the demonstrator use-cases**

In addition to the description of the users needs, WP5 decided also to collect information on existing tools and resources (training, expertises) that are available in the ELIXIR infrastructure in relation to the demonstrator use-cases to the knowledge of the partners.

<sup>3</sup> <https://zenodo.org/record/4672569#.YG8i0xRKj0o>

<sup>4</sup> <https://zenodo.org/record/4672588#.YG8IGRRKj0o>

The table to collect this information was developed building on the work done in M3.1 and D1.1: ELIXIR\_CONVERGE\_Demonstrator\_resources\_v1<sup>5</sup>

## 4.2 Method for collecting the data

- The user stories and scenarios with personas were developed by WP5 partners with initial guidance from WP2 and WP3 on the different categories of personas.
- The table on existing tools and resources was also filled by WP5 partners and target interviews of researchers for some demonstrator use-cases.
- Additional input was sought from the ELIXIR communities for demonstrator use-case 1, 3 and 4.

# 5. Results

## 5.1 The personas and their requirements

Users stories and scenarios have been developed for three use cases:

- Demonstrator use-case 1: Harmonised FAIR plant genotype & phenotype data management toolkit for Europe
- Demonstrator use-case 3: Common Data management plans for the marine metagenomics Community
- Demonstrator use-case 4: Federated access to human genomics data: GDPR

A scenario describes the needs of demonstrator use-case 6: FAIR organisation of biomolecular simulation information.

### 5.1.1. The data consumers

All the demonstrator use-cases are addressing data consumers who need to find specific categories of scientific data (e.g. human genomic data per cohort) and to be able to reuse it in new analyses. The common personas for use-cases are:

- the researchers and the students that have similar needs: finding all the relevant data, elements of quality about the datasets, possibility to sort it by relevant categories and reference tools for analysis. Note that the main persona of demonstrator use-case 6, which deals with modelling, is a consumer of public data.
- the data archive managers: make statistics about the use of their stewarded data to improve their services.

<sup>5</sup> <https://zenodo.org/record/4672600#.YG8m9BRKiQo>

Demonstrator use-case 1 cited also the trainers/teachers that have similar needs to the researchers and students to develop their training courses: reference datasets and pipelines and knowledge about the most relevant data archives.

The two last data consumers cited by demonstrator use-cases 1 and 3 are interested in the data management plan:

- the funders who want to measure their impact through statistics on the data set produced by the projects they funded and their citation or number of downloads. They also might be interested to make statistics on the DMPs: what are the shared practices, tools and resources cited.
- the managers of facilities or laboratories who want to help their users to comply with good practices in terms of data management and data analysis along the whole process.

### 5.1.2 The data producers

All demonstrator use-cases cite the researchers and post doc/PhD students as data producers. They have very similar needs: they want to know how to best collect their data and metadata, to store it along its life cycle, and finally where to publish it. In addition, researchers want to know how to formalize a DMP for their projects.

Demonstrator use-cases 1 and 3 cite the lab/facility managers that also want to implement the FAIR principles on the data they produce, e.g. genomic or phenotyping facilities, or to support and guide researchers for their needs in terms of data management and data analysis.

In case of sensitive data (Demonstrator use-case 4), cohorts owners have specific activities to deal with requests of data and Trusted Third Parties, for the use of sensitive data.

### 5.1.3 The workflow productors

This category is not included in Demonstrator use-case 1 and has different predominance in the other use cases.

- Demonstrator use-case 2: the researchers that create workflows want to store and document them to make them reusable by others. They also want to benchmark



them. The managers of bioinformatics facilities want to know what resources are needed to run the workflows they install for their users.

- Demonstrator use-case 4: the data stewards need workflows that ensure that privacy rules are fulfilled.
- Demonstrator use-case 6 is mainly about workflows (models). Its main persona, the modeller, needs a platform to store their models and link them to the data.

#### 5.1.4 Data owners/managers/controllers

This category is very important in Demonstrator use-case 4 in relation to sensitive human data. The data owners need to ensure that the GDPR rules are fulfilled and to measure the legal risks associated with their research projects. To make this possible, data managers as data custodians must ensure that the data is under strict control access along the whole data life cycle.

For the other use-cases, researchers need guidance to associate the right licence to their data and facilities/institutions on policies about data ownership and data licencing.

#### 5.1.5 Data curators

- All demonstrator use-cases cite the need for automatic checks of metadata (presence) and format compliance.
- Demonstrator use-case 4 must process the metadata in relation with its sensitivity.
- Demonstrator use-case 3 needs to develop quality checks on the data and the metadata.
- Up-to-date guidelines for properly maintaining and interconnecting metadata, e.g. to contribute towards the FAIRness of the data/metadata, are needed for all use-cases.

#### 5.1.6 Conclusions

The specifications of the demonstrator use-cases can be linked:

- to the biological material they deal with, its associated metadata capturing method, the knowledge about the environment in which it was sampled, specific identifiers and privacy management. Note that the three biological domains represented by Demonstrator use-cases 1, 2, and 4 are supported by ELIXIR communities.
- to the demonstrator use-cases which are generic, in terms of biological domains, that are dealing with very different types of data. Demonstrator use-case 2 is about a specific type of omics data, e.g. epitranscriptomics, and the associated analysis tools at the moment are part of the associated metadata while in demonstrator



use-case 6, the main data is the statistical model derived from the original data using advanced simulation techniques.

- to the phases of the data management plan they address (see table 1): all the demonstrator use-cases address the planning, preserving, sharing and reusing phases. However, there are other phases like analysing that are only relevant for some of them.

Table 1. Phases of a data management plan addressed by the ELIXIR-CONVERGE demonstrator use-cases (marked by an X).

Demonstrator use-case	Biological domain	Planning	Collect	Process	Analyse	Preserve	Share	Reuse
1	specific	X		X		X	X	X
2	generic	X	X	X		X	X	X
3	specific	X	X	X	X	X	X	X
4	specific	X		X		X	X	X
5	generic	X	X	X		X	X	X
6	generic	X			X	X	X	X

## 5.2 Expertise, resources and training: ELIXIR's support and gaps

To complete this first analysis, the WP5 partners have landscaped the current knowledge in terms of resources, training and expertise in relation with each demonstrator use-case and with each phase of a data management plan: (see ELIXIR\_CONVERGE\_Demonstrator\_resources\_v1)<sup>6</sup>

The resources were described either as generic or as dedicated to a specific biological domain. It is interesting to see that there are already a significant number of resources that can be leveraged to address the demonstrator use-cases. Many of them are already registered in ELIXIR recognized registries like bio.tools and FAIRsharing. Importantly, key resources are specifically distinguished as ELIXIR Core Data Resources and/or Deposition Resources so the scientific community can trust them for gaining access to the knowledge available there and/or deposit their biological data for archiving and further sharing and reuse. Some of those resources, e.g. EGA, allows a fine-grained control of who accesses

<sup>6</sup> <https://zenodo.org/record/4672600#.YG8m9BRKj0o>

what data. Some of the resources listed along the different demonstrator use-cases are part of the ELIXIR Node Service Delivery plans, which contribute towards the sustainability of them.

Some training courses and materials have been developed in particular for the management of sensitive human data, data management in general and standardization of plant phenotyping data. As we have seen in the work on the personas/users of the demonstrator use-cases, the researcher persona is always included in various positions, e.g., data consumer, data producer, data owner, PI writing a proposal, training. Therefore, we can say this category of persona will be particularly important. Also, a common terminology across the whole project is being developed, e.g. an extended and revised definition of personas, to allow generating fully coherent materials, best practices and guidelines.

Finally, a first round of identification of nodes' expertise needed to support the demonstrator use-case is available. This landscape is far from comprehensive and requires deep survey within ELIXIR nodes to move forward in its completion.

## 6. Conclusions

This is the first deliverable produced in the context of ELIXIR-CONVERGE WP5. As such, it has contributed to deepening the understanding of the pre-selected demonstrator use-cases. Selected use-cases represent a broad diversity of data types, with many associated challenges as well as bring together different sets of users roles, each of them with their own needs and expectations. The use of two differentiated waves of demonstrator use-cases has allowed us to rapidly move forward with some of them. On the contrary, the level of detail is at the moment differently among use-cases. As we progress in the project, and building in the initial experiences, we expect that those differences will progressively disappear.

The method implemented to describe the demonstrator use-cases in detail has been efficient to isolate the specific context of each case, and to identify specific tools/resources in relation with the use-cases that can be used for data management. Importantly, it has become clear that demonstrators will not cover all areas of the research data life cycle. This finding was not foreseen at the time of selecting them and is a direct product of the interactions with other WPs in the project.



The main aspect used to classify a new use case is the type of biological material linked to the data/metadata. Data/metadata is then linked to specific requirements and considerations, e.g. legal, use of ontologies, access mechanisms. The second is the relation of the personas with the research data associated to each demonstrator use-case, e.g. consumers/producers, and the type of data, including workflows, models, and the various types of experimental data.

The landscape analysis of the resources already available in relation with the demonstrator use-cases and the different stages of a data management plan showed that the guidelines developed by WP1 are very useful to clarify the scope of these stages before filling such a survey.

The personas/users related to the use-cases are very diverse but always include the researchers in various positions, e.g. data consumer, data producer, data owner, PI writing a proposal, and the data curators. This will be further developed in collaboration with WP2.

In close collaboration with WP3, we expect that demonstrator use-cases will provide feedback on the existing general knowledge model for research data management. Such feedback will set the foundations to propose specific extensions to the general knowledge model. Such extensions, potentially one per demonstrator, will allow researchers and data stewards and managers to have a greater level of detail when elaborating their data management plans.

## 7. Impact

The present work has allowed us to develop dedicated pages to the “Your domain” in the RDMKit website (<https://rdmkit.elixir-europe.org/>) for demonstrator use-case 1, 3, 4 and 6. In these pages, specific challenges linked with the demonstrator use-cases are highlighted, and dedicated resources to solve them are listed. These tools have also 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.

This work not only had an impact across ELIXIR-CONVERGE, it also allowed the different demonstrator use-cases to be aware of their needs in terms of data management and to be able to categorize them so that they will be able to get the correct support by experts from each expertise.





Demonstrator use-cases are connected in most of the cases with different research communities. Thus, demonstrators act as a proxy to gather feedback from researchers in those communities. The objective is to iteratively validate the outcomes associated with the demonstrator within ELIXIR-CONVERGE with the associated communities to ensure alignment and increase their mid/long term impact and usefulness of those efforts.

## 8. Next Steps

The perspective of the work is:

1. To complete the work with dedicated interviews of personas, in particular for the demonstrator use-cases of the second wave for which we have little data yet (Demonstrator cases 2 and 5). The aim will be to quickly identify what needs to be added to the ELIXIR RDMKit website in relation with these use-cases. For instance, Demonstrator use-case 5 contains specific aspects about human/IP protected data but the specific aspects dealing with such sensitive data have already been described. The complement should deal with the rest of the data.
2. To see how the Data Stewardship Wizard can be used to complement the RDMKit website and provide templates to deliver actual data management plans.

## 9. Deviation from Description of Action

3 month delay in the publication

