

D4.3: Service Uptake within Communities

Author(s)	Daan Broeder, Herman Stehouwer
Status	Final
Version	v1.0
Date	13/06/2016

Abstract: Deliverable D4.3 reports on the progress and achievements of WP4 in the first 12 months of EUDAT2020 with respect to task 4.2 “Service Uptake”. It describes the administrative infrastructure for the uptake of the EUDAT services by the research communities associated with EUDAT and the communities’ plans for the uptake processes. The communities include both core communities that have been collaborating with EUDAT in the longer term and communities that have become involved with EUDAT more recently via Data Pilot studies. This deliverable also contains an analysis of the proposals and plans resulting from the first EUDAT Call for Collaboration, and discusses feedback from research communities, including additional requirements for new functions and new or augmented EUDAT services.

Document identifier: EUDAT2020-DEL-WP4-D4.3	
Deliverable lead	CLARIN ERIC
Related work package	WP4
Author(s)	Daan Broeder, Herman Stehouwer
Contributor(s)	Research community uptake coordinators: Peter Danecek, Margareta Hellstrom, Steven Newhouse, Johannes Peterseil, Hannes Thiemann, Dieter van Uytvanck
Due date	31/02/2016
Actual submission date	13/06/2016
Reviewed by	Damien Lecarpentier, Carl Johan Håkansson
Approved by	PMO
Dissemination level	PUBLIC
Website	www.eudat.eu
Call	H2020-EINFRA-2014-2
Project Number	654065
Start date of Project	01/03/2015
Duration	36 months
License	Creative Commons CC-BY 4.0
Keywords	Research community, requirements, service uptake

Copyright notice: This work is licensed under the Creative Commons CC-BY 4.0 licence. To view a copy of this licence, visit <https://creativecommons.org/licenses/by/4.0>. 

Disclaimer: The content of the document herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services.

While the information contained in the document is believed to be accurate, the author(s) or any other participant in the EUDAT Consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the EUDAT Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the EUDAT Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

TABLE OF CONTENTS

EXECUTIVE SUMMARY	4
1. INTRODUCTION	5
2. THE INSTRUMENT OF COMMUNITY UPTAKE PLANS.....	7
3. THE UPTAKE PLANS IN THE CONTEXT OF OTHER EUDAT COMMUNITY INTERACTION	9
3.1. Relation to the Data Life Cycle Analysis by WP8	9
3.2. Relation to Other EUDAT Community Interactions.....	9
4. UPTAKE PLAN OF THE CLARIN COMMUNITY	10
5. UPTAKE PLAN OF THE LTER COMMUNITY	14
6. UPTAKE PLAN OF THE EPOS COMMUNITY	18
7. UPTAKE PLAN OF THE ICOS COMMUNITY	23
8. UPTAKE PLAN OF THE ENES COMMUNITY.....	27
9. UPTAKE PLAN OF THE ELIXIR COMMUNITY	30
10. UPTAKE OF EUDAT SERVICES BY THE DATA PILOTS.....	32
10.1. Data Pilot Requirements.....	32
11. CONCLUSIONS.....	36
ANNEX A. GLOSSARY.....	37

LIST OF FIGURES

Figure 1: The workflow between the EUDAT work packages and the research communities (Cn) for developing the Community Uptake Plans (CUP) and specifying service usage and service development requirements and the response in the form of a uptake plan response	8
Figure 2: The CLARIN resources and metadata infrastructure, types of resources, harvesting procedure and visualization in the CLARIN Virtual Language Observatory (VLO) metadata catalogue	11
Figure 3: LTER View on possible matching of LTER workflow with EUDAT services	15
Figure 4: LTER's DEIMS metadata tool	16
Figure 5: EPOS Organisation of Integrated Core Services and Thematic Core Services.....	18
Figure 6: EUDAT-EPOS-EIDA Relations	21
Figure 7: A schematic overview of the organization of the ICOS RI and the flow of data between the main nodes	24
Figure 8: Same as previous figure but now with EUDAT services indicated	24
Figure 9: ENES Data Infrastructure showing the relations between the different ESGF and IS-ENES Data Nodes and the different ENES related portals.....	27
Figure 10: The ELIXIR partners, the EMBL-EBI organization and the ELIXIR Nodes and Hub Structure.....	30
Figure 11: Combined Community (blue) and Data Pilot (red) service selections.....	33

EXECUTIVE SUMMARY

The EUDAT Collaborative Data Infrastructure (CDI) provides services for handling research data, primarily in the context of the European research landscape. Many of the EUDAT services have already been developed and are in the process of being taken up by research communities that are associated with EUDAT. This deliverable reports on the plans for the uptake of those services by the research communities. It also highlights further requirements that have arisen from the research communities in regard to the services, and which need to be taken into account in the development of the EUDAT CDI. This document presents an overview of the new or modified services that are being requested by the research community partners within EUDAT and also covers the services that are mentioned in the data pilot studies which were submitted as part of EUDAT's latest open call for collaboration.

The requests for existing and new services are discussed and put into the context of both the current EUDAT service roadmap (which is documented in deliverable D5.1 “Yearly reports on service building status and progress”) and the current landscape of data management services (which is partly documented in D4.2 “Data and Computing Landscape Characteristics and Scientific Communities’ Environment & Requirements”).

1. INTRODUCTION

The EUDAT project started in 2012 with the aim of developing a Collaborative Data Infrastructure (CDI) to provide data handling services for European researchers. The first phase of the project began by identifying a range of data managing services that would be of benefit to researchers in different disciplines, and then started work on developing the most generally useful of those services. The first phase of the project ended early in 2015 and was instrumental in developing preliminary versions of many of those initial EUDAT services. Now, in the second phase of the project, many of these services are in the process of being taken up by various research communities that are associated with EUDAT.

The process of taking up services like these across geographically dispersed research communities is not a trivial matter as the researchers from any particular community may be in different countries or cities (and therefore have to access and share a lot of their data remotely), and there may be different organisations or institutions in various locations that provide data storage for each specific community (which means the uptake of the services needs to be coordinated between several disparate entities). Consequently the uptake process involves considerable planning, both on the part of EUDAT and also by the relevant research communities and organisations.

To this end, EUDAT has worked with the research communities that are partners in the current EUDAT consortium, and also with research communities that have become involved with EUDAT through our latest call for Data Pilots. (The aim of these Data Pilot studies was to provide research communities with a large amount of data storage on EUDAT resources, plus support from EUDAT personnel to perform tests or do further development relating to one or more of the EUDAT services or to integrate EUDAT services into the community infrastructure.) Together, EUDAT and its research partner communities have developed concrete individual plans for the uptake of EUDAT services.

This deliverable reports on the progress with developing those uptake plans, and also includes plans and requirements for ‘new’ services or ‘adaptations’ to existing EUDAT services that have come to light during the discussions. (Ultimately, the Data Pilots should likewise result in (mini) uptake plans.) The uptake plans are specified in what are known as Community Uptake Plans (CUPs) – these are valuable as the main source of information about what needs to be done by the EUDAT enabling and operations teams in Work Package 6 (WP6) and by our service development team (in WP5) to help the research communities take full advantage of the EUDAT services.

The information about the uptake plans is presented in this deliverable in several parts covering (1) the nature and workflow of the CUPs and a summary of the contents of every CUP, as they are the main source of information for formalizing our service uptake processes and for determining requirements, (2) a summary of the Data Pilot uptake and requirements, and (3) an analysis of all of these, accompanied by contextual observations.

The information in this deliverable originated from the following sources:

- the different versions of the Community Uptake Plans (which were created through common endeavours by the research communities and EUDAT experts),
- discussions about the contents of the uptake plans that have been held between EUDAT and the core communities (which have been ongoing since the start of this phase of the EUDAT project) and between EUDAT and the communities that were awarded Data Pilots,
- presentations and discussions at the EUDAT Community Meetings and User Forums, and
- information from EUDAT WP8 relating to the analysis of mapping the uptake plans to the data life cycle.

It is important to be aware that much of the work that EUDAT and the research communities are doing is effectively cutting new turf. There are no magical pre-existing blueprints for the data services that research communities need, nor are there well established tried and tested methods for diverse research communities to take up new data handling services. This means that EUDAT and the research communities need to discuss matters, and then go away and consider things and perhaps test options, and then discuss further, and to

keep repeating that process to eventually reach a consensus about suitable approaches. In this regard, the discussions and presentations at the EUDAT User Forums and Community Meetings function as informal brainstorming sessions for generating creative solutions for further investigation. Those ideas are distilled with information from the other sources to then produce the Community Uptake Plans, which are regarded as the more formal baseline for the eventual collaborations.

Another point that must be borne in mind is that the latest EUDAT Call for Data Pilots closed quite recently and therefore EUDAT has only been in contact with the new research communities (that is, the ones that had not been involved with EUDAT prior to their involvement with the current set of Data Pilots) for a relatively short amount of time. Consequently EUDAT and those research communities have not yet clarified all the aspects of their Data Pilot plans, and hence there is still further work to do so that the plans are appropriately aligned with the EUDAT services.

Two of the research communities that EUDAT has been working with find themselves in somewhat exceptional circumstances in relation to the uptake plans.

- The VPH community, which is one of EUDAT's long-standing partners, does not have a specific CUP at all. The reason for this is that VPH does not currently have a community research infrastructure (RI) project which would fund and support the uptake of EUDAT services. Therefore it was agreed that EUDAT's contribution to VPH would consist of continuing the service uptake that was already started in the first phase of the EUDAT project.
- The CUP for the ELIXIR community is still in a very early phase – this is due to a recent redrafting of the earlier planned service uptake at the request of the community. While the process of formalising the ELIXIR CUP is underway, it will not be finalized until later.

Since this deliverable focuses on the CUPs, it contains many quotations from the different plans provided by the research communities working with EUDAT, however, for ease of reading, we have not specifically identified every quotation, but simply used the information from the research communities within the body of the text.

2. THE INSTRUMENT OF COMMUNITY UPTAKE PLANS

The task of working out and planning how a research community can and will take up an EUDAT service is not entirely straightforward, as was mentioned in the previous section. Therefore EUDAT and the research communities that are interested in using EUDAT services have been working together on developing Community Uptake Plans for each community. The intention is that these CUPs will specify what each research community expects from EUDAT as a service provider and also give details of how the relevant EUDAT services will be integrated into the community's own infrastructure. However, it takes some back-and-forth discussion between EUDAT and a research community to determine what is feasible from each side, so the successive incarnations of each CUP act as a basis for an on-going dialogue between EUDAT and the relevant community about the feasibility of the community's requests and visions regarding their use of the EUDAT services. In this way, the CUPs serve as the primary instruments for discussing the integration of EUDAT services into the research communities' infrastructures and for defining the roles that EUDAT and each community will play in the uptake process.

Uptake plans will be developed for all of EUDAT's collaborations with research communities that entail the use and/or development of data management services. The intricacy of each uptake plan depends on the complexity of the services that are required by a particular community. Obviously some of the requests that the research communities make are easier to address than others. For example, if a research community wishes to use EUDAT's generic B2SHARE service, this can be done in a straightforward way, so the CUP for that is relatively simple. However, when a community makes a request such as integrating a new community metadata profile into B2SHARE, there is more work involved in doing that, so discussions are needed to agree on a clear plan for the steps that will be undertaken to achieve that goal.

The success of each such plan relies on efficient communication and interaction between the research community partners and the EUDAT experts (mostly from WP4, WP5 and WP6) that will be involved in the uptake process. The CUPs are used as the basis for this interaction. The way this works is as follows.

EUDAT WP4 is responsible for Community Engagement and Requirements. WP4 liaises with the research communities (particularly with the community uptake coordinators) to determine what each community needs and wants in terms of EUDAT services – that information is then used, mainly by the research community uptake coordinators, to produce an initial CUP for each community. EUDAT and the research community can then further modify the CUP if it becomes necessary, though it is important to note that the community uptake coordinators are the primary owners of the plans and they each take responsibility for maintaining their community's plan. In addition, it has been agreed that for each CUP there will be a Community Uptake Plan Response (CUPR), which will be maintained by WP4. These CUPRs will provide details of what EUDAT will do in response to each CUP. The work to develop the CUPRs is currently in progress so, although the CUPRs are not finalized at this point, relevant information about the requirements from the research communities and suggestions for EUDAT's responses to those requirements are provided in the sections of this document that deal with the uptake plans for the individual research communities.

The way that the CUPs and CUPRs are used is as follows. EUDAT's WP5 (which is responsible for service development) has created service development templates (which may also include requests for modifications to the services). These templates describe the requests from the communities in detail. In contrast, the service creation and modification requests as they are recorded in the CUPs are from a higher level perspective, and hence do not go into details regarding the necessary development, but rather provide the context and background for the requests. EUDAT's WP4 then fills in the WP5 service development templates on the basis of the CUPs, with assistance from research community experts and WP5 service experts as needed. The filled-in templates are then added to the Service Development Requests List. If a research community has put in a request via a CUP that is for using existing services (without requiring any extra development work on EUDAT's part), then such requests are taken up by the service operations work package, WP6. This work package has teams of service 'enablers' who (in collaboration with the relevant research community) facilitate the provisioning of the service to the community. To do this, the WP6 enablers create highly detailed Data Project Plans that specify the configuration and resources involved with each of

the services that needs to be enabled. Despite this comprehensive planning process, it can happen that, in the course of their enabling work, the service enablers find that some modifications to a particular service are needed, in which case the enablers then also need to fill in a service development request template.

All of the requests that are on the Service Development Request list are subject to prioritization discussions in the EUDAT Technical Committee (TC). Decisions to accept or refuse a development request, along with the prioritization that is assigned and the underlying reasoning, are described in the CUPR and discussed with the relevant research community. This workflow is depicted in Figure 1.

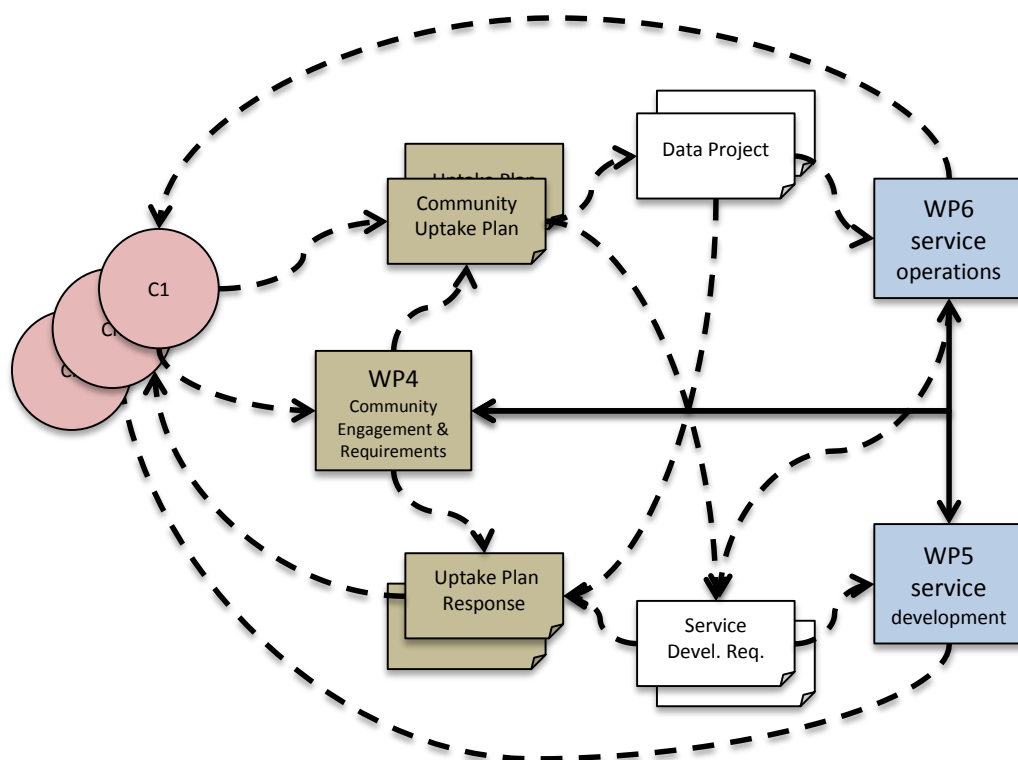


Figure 1: The workflow between the EUDAT work packages and the research communities (Cn) for developing the Community Uptake Plans (CUP) and specifying service usage and service development requirements and the response in the form of a uptake plan response

There is considerable variation between the research infrastructures of the different research communities currently involved with EUDAT, and this has a natural follow through effect on the uptake plans of each of the communities. The level of specificity of an uptake plan generally mirrors the current of stage development of the community's research infrastructure. Some research communities are still in the early stages of specifying the architecture for their infrastructure – so it can be hard for them to specify the services they need from EUDAT at this juncture, while other research communities already have an established infrastructure so they are in a position to give a clear statement of their requirements with respect to EUDAT's services.

Because EUDAT's goal is first and foremost to support and help European researchers to manage and share their data effectively, we strive to adapt to the research communities as their infrastructures mature. We have therefore agreed with the communities that the uptake plans will be regarded as being flexible in principle. This means that if a community decides that changes are needed, it will be possible to make changes to their CUP. Naturally making such changes would necessitate holding discussions on other levels about the amount of effort that would be required to implement the new requirements. Overall though, the uptake plans are primarily the responsibility of the research communities and should reflect their ideas and needs (rather than being instruments driven by EUDAT).

3. THE UPTAKE PLANS IN THE CONTEXT OF OTHER EUDAT COMMUNITY INTERACTION

3.1. Relation to the Data Life Cycle Analysis by WP8

EUDAT's WP8 has been analysing the extent to which specific new EUDAT services¹ and approaches are needed in order to augment the existing research communities' workflows – this WP8 analysis work should be regarded as an effort that zooms in on specific needs for specific workflows. WP4 efforts are broader identifying requirements beyond the scope of workflows or specific technologies. Nevertheless it follows that the requirements that are identified by WP4 and WP8 are largely overlapping.

The research community uptake plans are a reflection of the views of each individual community, and hence the WP8 findings about the usefulness of existing or new EUDAT services (which can be found in D8.1 "Report of Requirements") may differ from the analyses and plans of the research communities. Nevertheless it is highly useful for WP4 to use the D8.1 results to cross-check the existing uptake plans and to discuss them further, and hence those plans should be living documents mirroring the outcome of the ongoing discussions between EUDAT and the communities. The plans should also be based on discussions by other WPs within EUDAT.

Where it is appropriate, these types of suggestions have been included to lists and analyses contained in this deliverables, in the relevant context.

3.2. Relation to Other EUDAT Community Interactions

In addition to the discussions with the research communities about the uptake plans, EUDAT has two other vehicles of discussion and types of interaction with the communities:

- the EUDAT User Forums (the first of which was held in Rome, 3-4 February 2016) where EUDAT invites the EUDAT core communities and the research communities involved with Data Pilots to present their use cases and report on their progress with the uptake of the EUDAT services, and
- the EUDAT community meetings (the first of which was held in Lund, 12-13 January 2016, and hosted by the ICOS community, and the second in Utrecht on 12-13 May 2016, hosted by CLARIN) which are a new type of event that we have instigated to intensify the communication between the EUDAT boards (that is, the Project Management Board (PMB) and TC) and to keep them informed about and involved in EUDAT's daily discussions and developments.

At both the User Forums and the community meetings, the research communities present and discuss their plans – the contents of these presentations are then taken as valuable input to be considered in addition to the uptake plans. Such input has also been taken into account in the lists and analyses in this deliverable.

¹ These are new technologies that were identified in the first phase of EUDAT as the Generic Execution Framework (GEF), Dynamic Data and Semantic Services as B2NOTE.

4. UPTAKE PLAN OF THE CLARIN COMMUNITY

CLARIN Infrastructure Goals and Organization

CLARIN is the European research infrastructure for the social sciences and humanities. It aims to provide easy access to language resources (data) and language technology (LT services), not only to the linguistic research community but also to the wider humanities research communities, which could profit greatly from such technology (for instance, from different types of text mining). The CLARIN community has existed nominally since 2006 when it was first placed on the European Strategy Forum for Research Infrastructures (ESFRI) roadmap². The EC funded a CLARIN preparatory project that ran from 2008 till 2011 and which resulted in the CLARIN European Research Infrastructure Consortium (ERIC) being established as a coordinating body in 2011. On the current (2016) ESFRI roadmap, CLARIN has now been classified as a “Landmark”, designating that it has become a pillar of scientific excellence and competitiveness in the European Research Area (ERA). The CLARIN ERIC activities are funded by contributions from the ERIC member countries and by participation in EC projects. Currently the CLARIN ERIC has member countries that subsidize national CLARIN projects. The CLARIN infrastructure is based on CLARIN centres (of which there are currently 33) of different sizes and capabilities agreeing to provide access to language resources and technology in a standardized manner. There is no single centre that overshadows the others in importance and the CLARIN ERIC can only speak for the centres in relation to an agreed-upon and limited part of their activities. The national CLARIN projects and the CLARIN centres are represented in separate governance boards.

CLARIN in EUDAT

The CLARIN ERIC, as an independent entity, is a partner in the current phase of EUDAT, as is the EKUT partner that is operating a CLARIN centre, along with EUDAT data centre partners MPCDF and FZJ which have roles in providing general services beyond the general scope of EUDAT to CLARIN centres. In the first phase of the EUDAT project, CLARIN-affiliated organizations played, and now continue to play, important roles in developing the ideas for the EUDAT CDI.

CLARIN Technical Infrastructure

In addition to the CLARIN centres providing access to their language resources and technology, the CLARIN ERIC itself (or so-called CLARIN A centres) operate some central CLARIN services that are necessary for the functioning of the CLARIN infrastructure and that benefit the whole CLARIN community. CLARIN centres are certified for interoperability and are required to: (1) use Component Metadata Initiative (CMDI) type metadata for Language Resources and technology, which is a flexible component-based way to describe their resources and services, (2) use Handle-based PIDs for referring to resources, and (3) become members of the CLARIN Service Provider Federation and use SAML-based Federated Identity Management (FIM) for authenticating users who want access to restricted resources. As well as that the centres are required to be DSA-certified³ for their resource archiving workflow. The CLARIN infrastructure needs a few central registries, such as a CLARIN centre registry that stores information about the CLARIN centres as their OAI-PMH endpoint and also provides a number of central tools, such as a central metadata catalogue and a virtual collection registry (VCR).

CLARIN Infrastructure Requirements

Most CLARIN centres are also archives for long-term preservation of language resources that require data archiving services. Partly these are solved by local (institute-level) solutions or are taken care of by national services. But many centres have no fixed solution and are interested in services that EUDAT can offer. PIDs are either provided by running a local Handle PID service, or by using the EPIC service, as EUDAT recommends. With respect to the authentication of users and giving them access to protected resources, CLARIN has been very successful in establishing their own inter-federation, the CLARIN Service Provider Federation (SPF) that connects many European national identity federations with CLARIN resource providers. EUDAT should

² https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri-roadmap

³ <http://www.datasealofapproval.org/en/>

connect its B2ACCESS service to the CLARIN SPF. This would not only be beneficial to CLARIN but would enormously increase the potential user base of all the services that are connected to B2ACCESS.

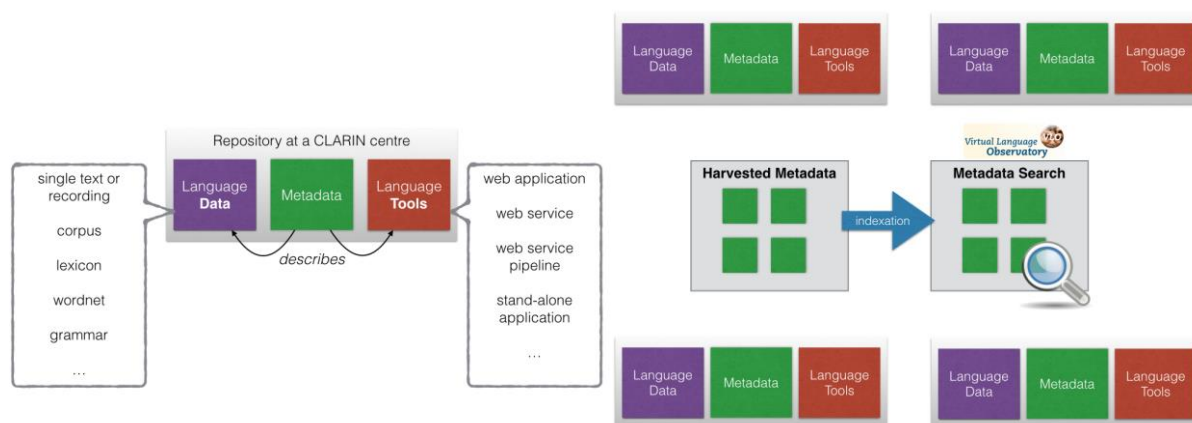


Figure 2: The CLARIN resources and metadata infrastructure, types of resources, harvesting procedure and visualization in the CLARIN Virtual Language Observatory (VLO) metadata catalogue

CLARIN Technology Opportunities

CLARIN is prepared to make some of its technology, particularly the CLARIN Virtual Collection Registry, available for use by other research communities.

In the past CLARIN has experimented with existing software for repository systems, such as DSpace and Fedora, and integrated these with B2SAFE functionality producing installable packages. The DSpace adaptation, originally developed by Charles University in Prague (UFAL), has gained a lot of popularity. It is available for CLARIN centres and others. If EUDAT did further development on this B2SAFE-enabled DSpace version, it would provide a huge improvement in the usability of B2SAFE.

Planned Uptake of EUDAT Services

For providing persistency for the data housed at the CLARIN centres, CLARIN has initiated a plan and procedure by which those centres can take advantage of expertise and resources provided by EUDAT and CLARIN to make use of the EUDAT B2SAFE service. The main aim is to connect the repositories of nine of the CLARIN centres to the EUDAT CDI for data sustainability. Connecting another three CLARIN centres (CLARIN-PL, The Language Bank of Finland (UHEL) and Talkbank-CMU) to the CDI will be scheduled after the initial nine have been satisfactorily connected. See Table 1 for further details about the planned uptake for the first nine centres

Centre	Country	Size (TB)	Technology prepared, e.g. iRODS installed,	Scheduled
SOAS, London University	UK	58	no	Feb-Apr 16
Språkbanken, University Gothenburg	S	10	no	Mar 16
CLARIN-AT, Austrian Academy	AT	5	no	Jun 16
Meertens, Dutch Royal Society	NL	12	no	Sept 16
CELR, Charles University	CZ	13	no	Nov 16
TLA, MPI for Psycholinguistics	NL	90	yes, v3	Feb-Apr 16
Finish Language Bank at UHEL, Finish CLARIN Centre	FL	?	no	tbd
Talkbank project from Carnegie Mellon University	USA	?	no	tbd
CLARIN PL centre, Wroclaw University	PL	?	No	tbd

Table 1: The CLARIN B2SAFE Uptake Planning

To prepare the uptake of B2SAFE by CLARIN centres, EUDAT and CLARIN organized a special training event for the staff at the CLARIN centres. Participation of centre staff in the event was one of the prerequisites for CLARIN centres to be selected in the first cohort of centres where B2SAFE would be enabled.

CLARIN's choice of using B2SAFE for persistency of data was based on earlier experiences of having successfully connecting a number of CLARIN community repositories using the B2SAFE service. Currently B2SAFE is being used by the TLA, EKUT and LINDAT CLARIN centres (see deliverable D6.1 "Report on Status and Progress of Operations"). Such positive experiences are highly influential in the decision making process when it comes to deciding to take up a specific complex technology, like B2SAFE.

Some interesting details in relation to this are that CMU (the Talkbank repository at Carnegie-Mellon University), which is actually a non-EU institute, will have an interesting consequence of its integration with B2ACCESS – namely that our AAI coverage will be expanded outside the EU, and secondly that CLARIN can encourage and promote the use of EUDAT services through the CLARIN RI, in contrast to connecting to direct local services (as in the case of The Language Bank of Finland that is hosted by the CSC data centre EUDAT partner).

CLARIN Service Development Requirements

To broaden the possible uptake of CLARIN services, CLARIN suggests additions to the following EUDAT services.

B2DROP-based workspace services: In the context of the development of their language technology-processing CLARIN is interested in having (language technology) web-services be able to access B2DROP-based workspaces (including taking care of access restrictions), which would also require that this process be interoperable with B2ACCESS and that B2ACCESS be connected to the CLARIN SPF. Basing such workspaces on the B2DROP service would make provision for storage and sharing, but also make it possible to easily deposit workflow results in B2SHARE.

For this to happen, several things would be required.

- User Delegation from B2ACCESS to the services would be needed. There was a successful experiment by CLARIN using ownCloud (the technology basis of B2DROP) and Unity IDM (the technology basis of B2ACCESS), as a result of which CLARIN expects that the technology will not be a limiting factor.
- In addition to catering for SAML-based IdPs, X.509 certificates⁴ and OpenID⁵-based Identity Providers, B2ACCESS would also need to be able to accept identities from a server-side LDAP⁶ interface. This would make it possible to use B2DROP from accounts in the CLARIN IdP. The technology for this is currently being tested within CLARIN.

Computational services: CLARIN has expressed interest in the possibility of having EUDAT service providers hosting services, which are currently available as B2HOST. CLARIN has offered to work with EUDAT to discuss what would constitute an attractive proposition for the research communities.

Access to EUDAT services: CLARIN would like to connect to B2ACCESS in such a way that users with an account at the CLARIN IdP (the CLARIN homeless researchers) can have access to EUDAT B2ACCESS (enabled) services.

The General Execution Framework (GEF): GEF was the research subject of an EUDAT working group active in the first phase of the EUDAT project. The activity has now been taken up by the current phase of EUDAT in the JRA WP8 and by the service development team in WP5. CLARIN suggests having CLARIN WebLicht workflows executed on the Generic Execution Framework (GEF). (The use case has not yet been clearly specified.)

The above subjects are now being put on the EUDAT service development requirements list and priorities will be discussed by the EUDAT technical boards and service area managers.

⁴ <https://en.wikipedia.org/wiki/X.509>

⁵ <https://en.wikipedia.org/wiki/OpenID>

⁶ https://en.wikipedia.org/wiki/Lightweight_Directory_Access_Protocol

5. UPTAKE PLAN OF THE LTER COMMUNITY

LTER Infrastructure Goals and Organization

The LTER Europe community⁷ (Long Term Ecological Research Network) is a network comprising around 455 long-term research sites (420 LTER sites and 35 LTSE Platforms) in 22 national LTER networks in Europe. Long-Term Ecosystem Research (LTER) is an essential component of worldwide efforts to better understand ecosystems. This comprises their structure, functions, and long-term response to environmental, societal and economic drivers. LTER contributes to the knowledge base informing policy and to the development of management options in response to the Grand Challenges under Global Change. LTER was considered not yet completely ready for inclusion on the 2016 ESFRI roadmap but was classified as an emerging project and considered to be scientifically excellent and working in a strategic area. The European LTER networks are currently funded as an Integrated Activity project eLTER 2020 that will help advance their development. LTER participates in the ENVRIplus cluster.

LTER in EUDAT

LTER is represented in EUDAT by the partners EAA (Environment Agency Austria, Austria), UNS the BioSense Institute at the University of Novi Sad, Serbia and the Research Centre Juelich (Germany) that already provides IT services to the LTER community in Germany (e.g. TERENO). EAA already participated in many EUDAT activities during the first phase of EUDAT. With their expertise in semantics and ontologies LTER is active in the EUDAT Semantics Working Group and in WP8's work on developing semantic technology.

LTER Technical Infrastructure

LTER Europe (by the means of eLTER and EUDAT) aims to provide a frame for an integrated discovery and access to metadata resulting from the distributed network of data providers within the LTER Europe network. The Data Integration Platform (DIP) is planned as a “one-stop-shop” interface for an end users (either researchers or experts) integrating metadata from the different distributed data sources (that is, the data providers/nodes).

Those data provider/nodes can either provide:

- a) existing service endpoints (in terms of metadata end point, data storage and services; advanced capabilities; e.g. TERENO) or
- b) deployment of standard virtual nodes (VN) providing regular or basic capabilities, which will be developed during the H2020 eLTER project.

Communication and information transfer between the data nodes and the central discovery portal is done via standardized service interfaces (e.g. OGC CSW for metadata).

This logical architecture aims to provide a scalable solution for data management for the LTER network. So, different data providers/nodes can also be registered in other networks, e.g. European or national sectorial portals. The Data Integration Platform will provide the central node to link LTER data into other networks (e.g. DataOne, EUDAT, GEOSS, etc.).

Open points which still needs to be implemented for the LTER data management framework (unless they are already provided by advanced data nodes) are:

- a) persistent identifiers (PID) for static (e.g. file-based) and dynamic data series which will be tackled by the use and inclusion of B2HANDLE used in B2SHARE,
- b) handling of user authentication (AAI) in the communication between DEIMS as metadata editor and B2SHARE as a data repository by the use of B2ACCESS, and
- c) tracking the provenance of the data (especially when data aggregation, assimilation and gap filling is applied) by the use of semantic annotation will be evaluated

⁷ <http://www.lter-europe.net/>

For these aspects developments are also undertaken within the related H2020 project eLTER. The LTER Europe data integration frameworks aim to implement standards wherever possible, namely the ISO19115/19139 and Environmental Markup Language (EML) for dataset metadata, INSPIRE Environmental Monitoring Facilities (EF) for research site documentation and exchange, and OGC SOS for the standardized exchange of time series data.

In Figure 3 there is a schematic representation of the LTER dataflow together with the B2-Services that could play a role. This workflow shows many possible uses of EUDAT services.

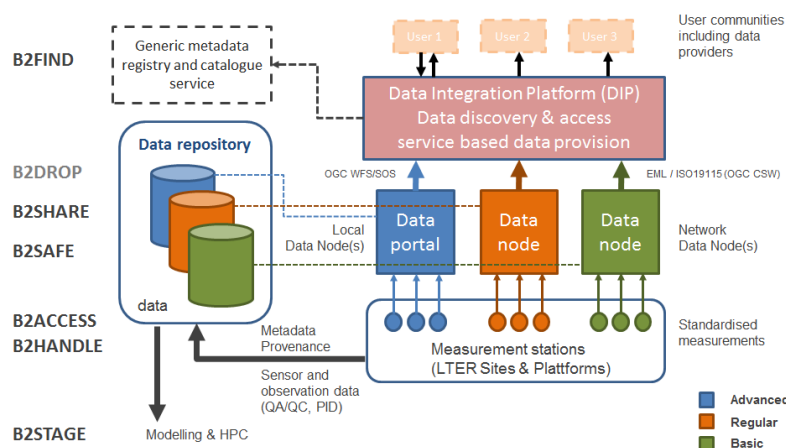


Figure 3: LTER View on possible matching of LTER workflow with EUDAT services

One main activity is integrating the DEIMS Research Site and Dataset Registry⁸ and metadata editor with B2SHARE so that users can directly deposit data sets into B2SHARE while providing metadata, using the B2SHARE API. The basis for this integration is DEIMS, a Drupal-based tool developed by the LTER community to edit and search in local metadata. DEIMS will become an important integration element with EUDAT services focusing on environmental data. Figure 4 illustrates this use case.

⁸ See <http://data.lter-europe.net/deims/>

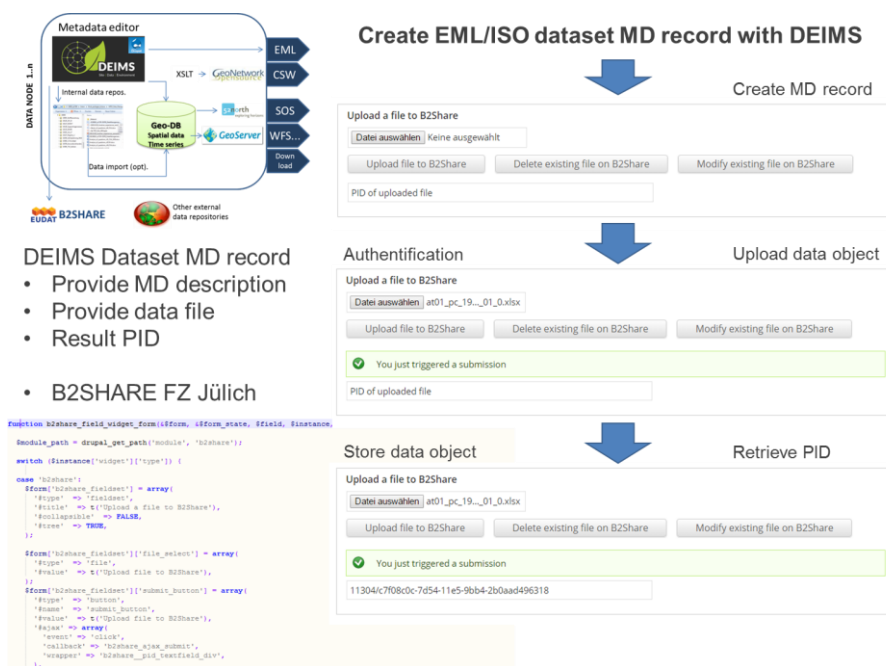


Figure 4: LTER's DEIMS metadata tool

LTER Technology Opportunities

The DEIMS tool could also be used by other communities that want to deposit metadata and observation data into B2SHARE. DEIMS is an open source development led by ILTER, US LTER and LTER Europe. LTER is prepared to discuss sharing this tool with other communities.

LTER Uptake of EUDAT Services

B2SHARE: As already discussed previously, a main activity in the uptake of the EUDAT services will be the integration of DEIMS and B2SHARE as a repository for file-based data using B2HANDLE as a PID service.

B2HANDLE: Handles or DOIs become available through the use of the DEIMS/B2SHARE connection for archived data sets. The extension of the B2HANDLE service to provide PIDs for data services would be a useful extension within LTER Europe.

B2FIND: LTER Europe aims to make the metadata on datasets and research sites discoverable in B2FIND. Furthermore, LTER Europe would also want to use B2FIND technology, using the B2FIND source code to implement (with some expertise) an LTER community discovery portal. EUDAT could also consider delivering B2FIND as a service, running a separate B2FIND instance for the LTER community.

B2SAFE: Based on the requirements analysis mentioned in the D8.1 deliverable, B2SAFE will be evaluated as a possible solution to ensure data persistency and replication, which is currently one of the issues to be solved for the central data nodes (cDN). Alternatively the use of B2SAFE as a distribution service for virtual machines for the virtual nodes is being considered, but needs further technical evaluation of the options. This can only be done together with EUDAT developers and experts.

B2STAGE: is considered in a separate data modelling use case where data from the DIP is staged to HPC centres. This is mentioned in the Uptake Plan as an additional use case and has to be worked out further and will not be tackled in the next implementation steps.

LTER Service Development Requirements

B2NOTE: The B2NOTE service concept is, in fact, a complex of services for the semantic annotation of data that was a study object in the semantic WG during the first phase of EUDAT, and it has been taken up as a JRA activity within this phase of the project. LTER proposes to use B2NOTE in the metadata-editing workflow, i.e. in DEIMS. Annotations should be stored in an EUDAT RDF triple store (prematurely referred to as

B2TRIPLE). This should not only make it possible to annotate Data Objects (e.g. data files) with keywords, but furthermore allow users to annotate elements (e.g. columns) with concepts from an controlled vocabulary or ontology. Semantic annotation is also the subject of the Data Pilot “B2ANNO” issued by a related community. More precise requirements are available in the LTER Uptake Plan and WP8 documents. LTER staff is participating in the relevant activities of the JRA WP8 and the Semantics working group.

B2SHARE & Environmental Thesaurus: There is need for the integration of an Environmental Sciences thesaurus (e.g. SKOS compliant controlled vocabulary like EnvThes⁹) in B2SHARE, either directly or through connecting B2SHARE to the EUDAT triple store and allowing it to choose vocabularies. This is a prototype implementation of the Ontology Lookup Service now specified as part of the WP8 activities. The resulting metadata should contain a link and description of the vocabulary used for metadata creation.

B2ACCESS: The user authentication request for the EUDAT CDI (e.g. accessing B2SHARE) should be passed from DEIMS using B2ACCESS for data upload and download. The mapping of the local DEIMS user to user roles in the EUDAT CDI is the subject of further evaluation and is a requirement for the LTER Europe Data Infrastructure.

Computational Cloud Service or Hosting Service: LTER suggests having a EUDAT computational cloud facility that will allow them to deploy the LTER Virtual Nodes (VN) Virtual Machines. The deployed VNs should be available over longer periods.

The above subjects are now being put on the service development requirements list and priorities will be discussed by the technical boards and service area managers.

⁹ See <http://vocabs.ceh.ac.uk/evn/tbl/envthes.evn>

6. UPTAKE PLAN OF THE EPOS COMMUNITY

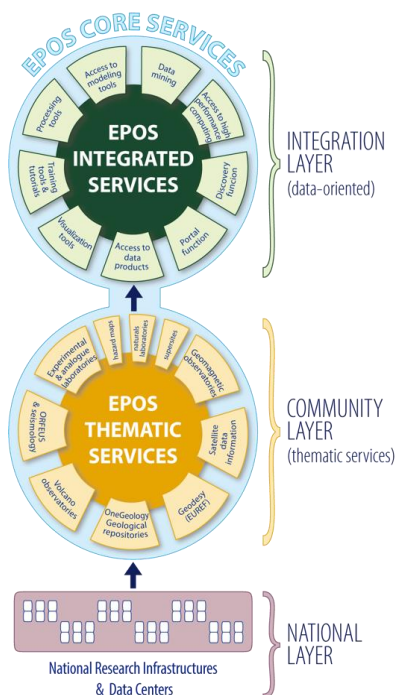
EPOS Community Goals and Organization

The *European Plate Observing System* (EPOS)¹⁰ is the integrated solid Earth Sciences research infrastructure. It searches to create a pan-European infrastructure to monitor and unravel the dynamic and complex solid Earth System. EPOS was approved by the ESFRI and included in the ESFRI roadmap in December 2008. EPOS is a long-term integration plan of existing national and international research infrastructures and was designed to integrate data and services from different Solid Earth scientific communities, like seismology, geodesy, volcanology, etc.

The EPOS community is organized in the following parts:

1. a national layer composed of National Research Infrastructures (NRIs) providing data and services;
2. the community layer, which is grouped into Thematic Core Services (TCS) and composed of pan-European e-Infrastructures – these disseminate data and services of a single discipline provided by NRIs, e.g. TCS Seismology is offering among others the federated service ORFEUS/EIDA*¹¹, and
3. an integration Layer, known as the Integrated Core Services (ICS) which is the e-Infrastructure designed and maintained by EPOS – this is the place where the integration of data and services provided by the TCS occurs.

*ORFEUS/EIDA (European Integrated Data Archive) is a well-established example of a Research Infrastructures providing data and services, and is one component of the TCS Seismology with dedicated governance. It has been serving the seismological community for many years providing access and management functionalities for seismic waveform data and related metadata. It complies with globally accepted standards.



EPOS Integrated Core Services (ICS)

provide simplified access to multidisciplinary data and data-derived products, combine data with modelling results (simulations), processing and visualization tools

Thematic Core Services (TCS)

community-driven infrastructures provide discipline-specific data services, these will build on pre-existing international collaboration/organizations (e.g. ORFEUS). The various communities organize their services. Seismology (**EPOS-S**) will provide and extend waveform data offerings through **ORFEUS/EIDA**.

National Research Infrastructures and facilities

provide services at national level and send data to the European thematic data infrastructures.

Figure 5: EPOS Organisation of Integrated Core Services and Thematic Core Services

¹⁰ <http://www.epos-eu.org>

¹¹ <http://www.orfeus-eu.org/eida/>

EPOS in EUDAT

The EUDAT partners that represent the EPOS community in EUDAT are INGV, KNMI and GFZ that all have a relevant role in EPOS and operate an ORFEUS/EIDA primary node. The EUDAT data centre partners supporting the EPOS community are CINECA, SARA and KIT. They provide the required storage and computational resources and replicate data from the three ORFEUS/EIDA pilot nodes, INGV, KNMI and GFZ, respectively.

The EPOS Technical Infrastructure

Following the organizational structure described above, the EPOS architecture is basically composed of three fundamental layers – ICS, TCS and NRIs.

The ICS layer represents the EPOS novel e-infrastructure consisting of services that will provide access to multidisciplinary data and products to different stakeholders inside and outside the scientific community. Products and services include data, synthetic data from simulations, processed data products, as well as processing and visualization tools. This key element of the ICS in EPOS will provide a single point of access where users can discover and access data, data products and services available through TCS and NRIs. A set of integrated services for multidisciplinary analysis data and processing may be delegated to external resources. The technical interface between TCS and ICS is the compatibility layer, which guarantees communication and interoperability. The ICS e-infrastructure will include the EPOS portal and its key functions: the Application Programming Interface (API), the metadata catalogue, the system manager and the services that will support data discovery, interactions with users as well as access to and integration of data.

The TCS layer is mainly formed by single communities, which will be integrated into the upper layer (ICS). In practice, the TCS constitutes the community-specific integration (e.g., in seismology, volcanology or geodesy). They represent a governance framework where each community discusses their specific implementation, best practices and sustainability strategies as well as legal and ethical issues.

The NRI layer is formed from a large variety of national research infrastructures, ranging from large geographically distributed sensor networks to very specialised research facilities or laboratories. They are managed on the national level and integrated into community-specific TCS.

ORFEUS/EIDA is one of the most advanced services of TCS Seismology and of EPOS in general, providing transparent access to the data archived in a federated structure of several European seismological data centres. Data from the distributed archives are currently accessible via the ArcLink proprietary protocol developed by GFZ, and more recently, a community agreed standard API (International federated of Digital Seismograph Networks, FDSN), based on web services.

ORFEUS/EIDA today is formed by ten primary (and two secondary) European data centres, which collect and archive data from seismic networks deploying broadband sensors, short period sensors, accelerometers, infrasound sensors and other geophysical instruments. In EUDAT three data centres of this federation are participating as partners. EIDA operations could be improved by the employment of various EUDAT services (see the B2SAFE, B2FIND, B2SHARE, B2STAGE descriptions in the following subsection).

The seismological community has well-established standards for the exchange of data (i.e. format and services) developed in the framework of the Federation of Digital Seismograph Networks (FDSN). These standards were adopted by EIDA. Further standards and services are in collaborative development within the seismological community.

EPOS Uptake of EUDAT Services

B2SAFE: The EUDAT data centre partners supporting the EPOS community are CINECA, SARA and KIT. They provide the required storage and computational resources and replicate data from INGV, KNMI and GFZ, respectively.

- INGV has hosted an installation of B2SAFE since the previous phase of the EUDAT project. Data from the continuous waveform archive amounts to about 60TB and is synchronized daily to CINECA. Updates to newer versions of B2SAFE as well as homogenization of the configurations are planned.
- KNMI currently hosts an installation of B2SAFE configured to replicate the ORFEUS continuous waveform archive to SURFsara. The size of the archive is about 50TB and data up to 2014 have been replicated. The current installation is based on iRODS 3.3 and the configuration is slightly different from the one operated at INGV in that here PIDs are generated at both sides (KNMI and SURFsara) and then linked. Further work is required to automate the replication mechanism. Updates to newer versions of B2SAFE as well as homogenisation of the configurations are planned.
- GFZ is currently exploring the possibilities for installing and operating B2SAFE in order to replicate its archive of continuous waveforms (80 TB) to the facilities offered by KIT. GFZ is testing the newest development version of B2SAFE. Deployment was completed and operational by the end of September 2015 (M06). Alignment of the versions and configurations with INGV and KNMI are continuously monitored and discussed.

The deliverable D6.1 “Report on Status and Progress of Operations” describes the current implementation status and the use of B2SAFE at INGV, KNMI and GFZ in more detail.

B2ACCESS: EPOS, in particular at the ICS level, requires a mechanism to securely provide federated identity management and accounting within a federation. It should be interoperable with the major existing mechanisms, be flexible and facilitate usability by providing different entry levels, e.g. from simple user/password to SAML and X.509 certificates.

On the TCS level, EIDA is developing and testing an authentication service that is able to contact different Identity Providers (IdPs) to authenticate the user and issue a digitally signed token. This token should be forwarded to any internal service in order to access restricted data. This authentication service is being designed to be compatible with most of the existing (de facto) standards.

Due to the anticipated number of users no big load is expected. Overall the speed of the service is however considered to be essential as EPOS users rely heavily on synchronous services (e.g. FDSN web services).

B2STAGE/GEF: EPOS, like any other complex e-infrastructure, seeks to provide data and data-derived products. Products result from data analysis of various kinds or from simulations. They can be part of the e-infrastructure and be provided as services, or they can be the result of workflows created by individual researchers. Some of these processes are CPU-intensive or rely on specialised platforms, and therefore require data staging onto such computational resources.

Moreover, the availability of workflow support when manipulating and processing data is essential in order to ensure reproducibility. The generic execution framework (GEF) solution provided by EUDAT might become a relevant component that EPOS would rely on to perform rather sophisticated and CPU-intensive analysis workflows.

Users or integrated EPOS services needs to stage Orfeus/EIDA data from the B2SAFE repositories to some HPC facilities and vice versa in a convenient way. Taking the data directly from the EUDAT data centre’s B2SAFE repository to a potentially nearby HPC facility would minimize data movement to and from the users or services. The GEF provides the opportunity to offer data pre-processed with relatively simple procedures or micro-services close to the B2SAFE storage thereby reducing the amount of data served by EIDA. Examples of such processing in seismology are filtering and down-sampling of seismograms, or rotating of components (e.g. radial vs. transversal) and delivering only selected ones.

The various EPOS (sub-)communities organize their own thematic services (TCS) and Seismology (**EPOS-S**) will provide and extend waveform data offerings through **ORFEUS/EIDA** the European Integrated Data Archive,

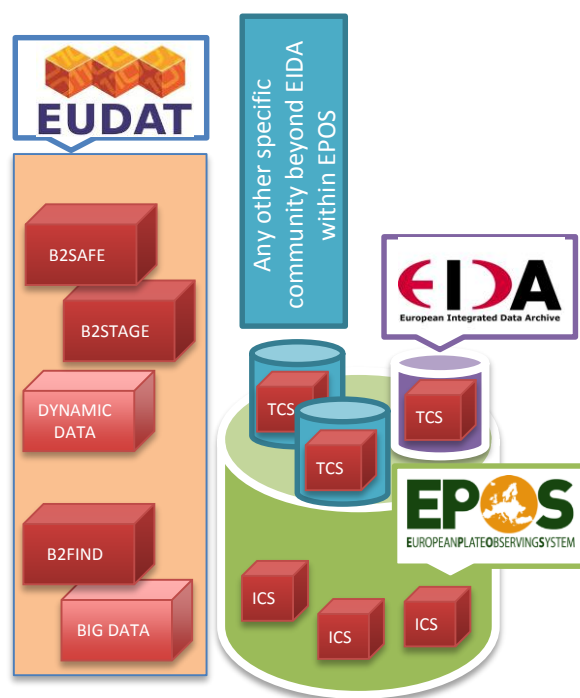


Figure 6: EUDAT-EPOS-EIDA Relations

which is a federated seismological data centre. EIDA was selected to collaborate and represent EPOS in EUDAT because of its size, culture in data sharing, existing widely accepted standards for data formats and exchange protocols, along with its well-defined development plan. See **Error! Reference source not found..**

EPOS Service Development Requirements

B2SAFE extension /access to replicated data: In addition to using the existing B2SAFE service for data persistency, EPOS has requested modifications and extensions of B2SAFE with respect to the following.

- Access for scientific or service purposes to DO copies housed at other EUDAT nodes than the original insertion node – This includes a service providing the location of the optimal (closest) copy but using location-independent PIDs, as opposed to identifying the physical location of DOs and resolving to the optimal location, determined by a set of metrics. Access protocols should include HTTP/HTTPS. Control of access is required for restricted data.
- Extended data management capabilities of B2SAFE providing services to manage distributed and federated datasets with multiple copies – Anticipated management tasks for this include: synchronization and verification of replicas, versioning of data, remote merging of versions, mapping of federated data, generation and association of metadata to DOs.

B2DROP extension / integration with B2SAFE: Another interesting use case for EPOS is when a researcher repeatedly requests a selection of EIDA data to work on in their research areas, modifying only a few parameters at a time until the desired dataset is obtained. It is very useful for the user to have these selections and the results from previous requests synchronized in their own cloud space. When this data is available in a B2SAFE instance, there should be an option to *redirect* the result of the request to a personal B2DROP account, instead of downloading the data.

The EPOS community would like EUDAT to consider the integration of B2SAFE with B2DROP, especially with regard to using B2DROP as a way to share ‘references’ to data deposited in B2SAFE. Such a feature would allow users to have virtually unlimited storage space in their B2DROP based workspaces, i.e. storing a collection of references would take next to no space. The EPOS use case description speaks of Virtual Digital Objects (VDO). A collection of such pointers could be stored in a ‘virtual file’ that would then be shared with others, and when a user needed to actually work on them, they would just need to instantiate the data by downloading the data into the B2DROP-based workspace.

B2FIND extension / metadata in the CDI: Another requirement from EPOS is to be able to use the B2FIND service (or an alternative catalogue) as a catalogue for the data deposited with B2SAFE. This implies the possibility to deposit metadata in B2SAFE. The metadata catalogue should provide PIDs that make it possible to download and stage data. The last point requires B2STAGE to work with PIDs and for B2FIND to be connected to all metadata in the CDI; the metadata catalogue should be federated and also include a newly-established EPOS catalogue.

The requirements with regard to separate metadata deposition have already been intensively discussed within EUDAT and have high priority.

B2ACCESS / AAA for EPOS specific services: EPOS, in particular at the EPOS Integrated Core Services (ICS) level, requires a mechanism to securely provide federated identity management and accounting within a federation. It should be interoperable with the major existing mechanisms, be flexible and facilitate usability by providing different entry levels, e.g. from simple user/password to SAML and X.509 certificates.

On the Thematic Core Services (TCS) level, EIDA is developing and testing an authentication service that is able to contact different Identity Providers (IdPs) to authenticate the user and issue a digitally signed token. This token should be forwarded to any internal service in order to access restricted data. This authentication service is being designed to be compatible with most of the existing (de facto) standards.

To accommodate the EPOS AAA requirements, we are investigating whether the EUDAT B2ACCESS service and technology can be leveraged. Three options are conceivable: (a) connecting EPOS Services to B2ACCESS, (b) providing EPOS with an instance of B2ACCESS that is deployed/maintained by EUDAT or (c) providing EPOS with the technology and expertise so it can run its own instance. For the first two options to be viable, one of the EUDAT centres would need to address various legal issues in relation to user privacy and sustainability issues so that transport and storage of privacy related user information is covered. With respect to the second option, there is an additional issue that needs to be considered, namely whether such a method of service provisioning is suitable as a business model for EUDAT (since up till now EUDAT has been primarily providing services to all its research community customers, whereas in this case it would be providing dedicated services to specific communities).

7. UPTAKE PLAN OF THE ICOS COMMUNITY

ICOS Infrastructure Goals and Organization

ICOS (Integrated Carbon Observation System) is a pan-European Research Infrastructure for quantifying and understanding the greenhouse gas balance of the European continent and adjacent regions. ICOS RI has a long-term commitment to providing standardized, state-of-the-art observational data on greenhouse gases and ancillary parameters to all interested parties. The RI became an ERIC in November 2015, and is now in its implementation phase.

ICOS RI is a highly distributed infrastructure, the Head Office is located in Helsinki, Finland and the Carbon Portal (CP), which is the community's data centre, is operated by Sweden and the Netherlands and located in Lund, Sweden.

ICOS relies on observation stations, which can be found in Belgium, Finland, France, Germany, Italy, the Netherlands, Norway, Switzerland and Sweden. Expert centres (one each for Atmospheric, Ecosystem and Ocean themed data) perform quality control and aggregation of the sensor data from the stations, after which ICOS data products are sent to the CP for curation and storage. The CP will also curate externally produced “elaborated” data products, typically spatial and temporal modelled maps of greenhouse gas emissions.

ICOS in EUDAT

The ICOS community is represented in EUDAT by the ICOS ERIC (Finland) and the University of Lund which hosts the ICOS Carbon Portal. The EUDAT partner research centres Juelich (JFZ) and the CSC-IT data centre were already actively providing services for the national ICOS organisations.

ICOS Technical Infrastructure

The Carbon Portal is responsible for organizing and coordinating all data management issues in ICOS (see Figure 7). It will build up and maintain a metadata database for all ICOS data objects based on semantic technology. (The EUDAT semantic service B2NOTE and the triple-store technologies EUDAT is considering in the JRA work packages could be used here.) This database will be the backbone of the CP's own data discovery and visualization tools, and will also be used for interfacing with other metadata portals and cataloguing services. (EUDAT's metadata catalogue B2FIND is one of these.) The CP will keep local copies of most (non-raw) data objects, but will rely on an external storage facility for safe, sustainable long-term archiving of the ICOS data (B2SAFE). In addition, making larger datasets (aggregations of ICOS data with other data including climatological and meteorological information) available to external users for HPC processing is also envisaged (B2STAGE and interfacing with e.g. EGI and other HPC and HTC service providers).

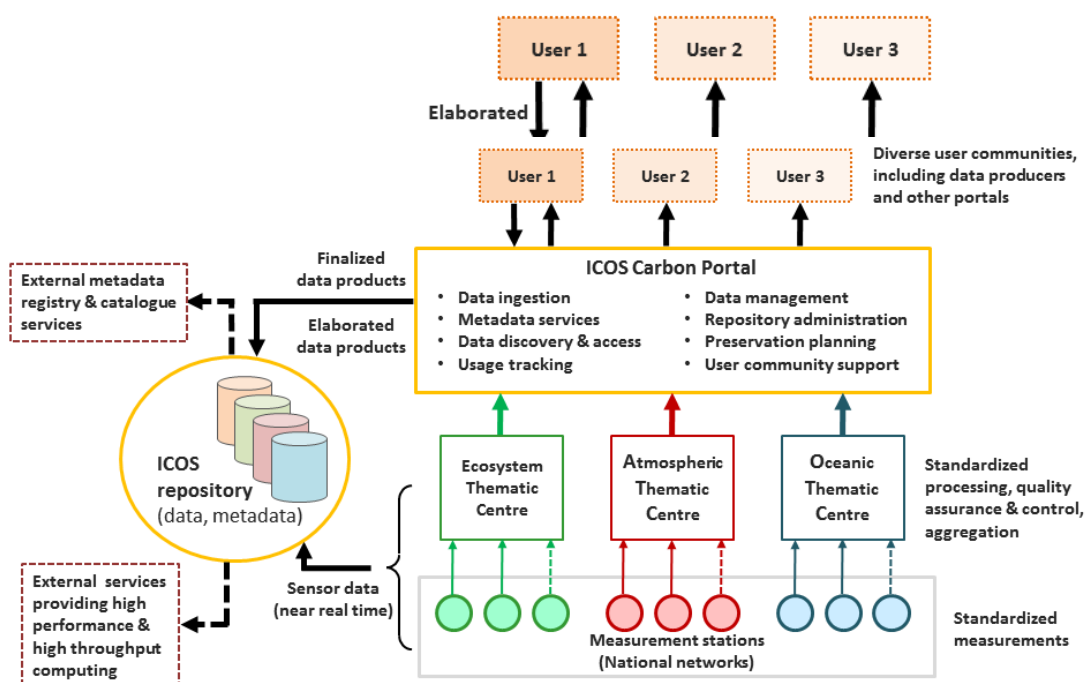


Figure 7: A schematic overview of the organization of the ICOS RI and the flow of data between the main nodes

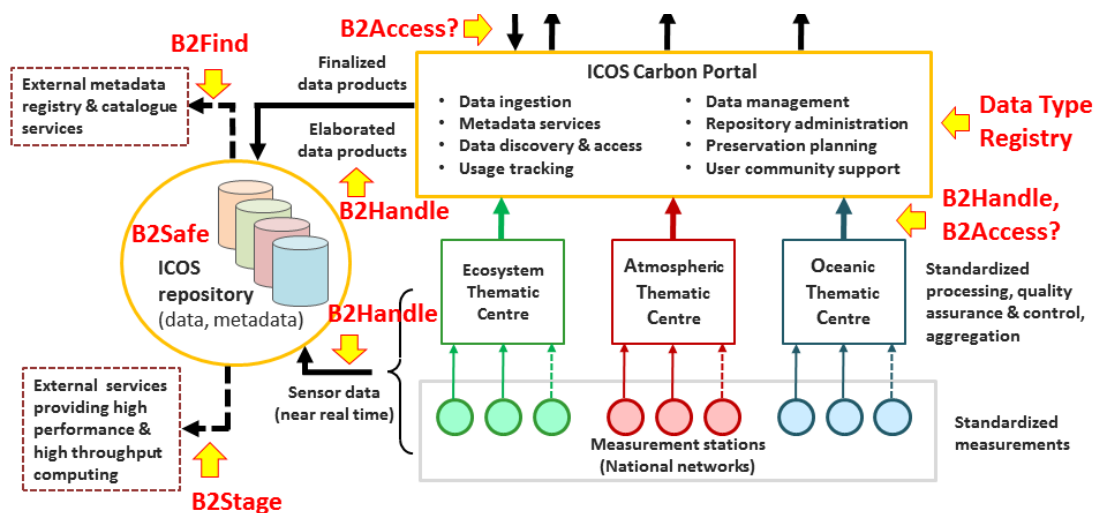


Figure 8: Same as previous figure but now with EUDAT services indicated

Note that during the start-up phase of ICOS (2016), the direct transfer of raw sensor data collected at the measurement stations to the repository will not be available. Instead, these data objects will be passed via the Thematic Centres to the Carbon Portal, which will ingest the sensor data into the repository in a similar manner to the finalized data. This will allow adequate time for development and testing of all required software components and interfaces, such as the one between the Carbon Portal and the B2SAFE service that forms the back end of the ICOS repository.

ICOS Infrastructure Requirements

In the current plans, the ICOS technical infrastructure will aggregate the data from the measurement stations to the CP without the involvement of external data management service providers, except for using a PID service to identify all generated data. Externally provided services for persistent archiving, publication (DOIs), discovery, access and (semantic) annotation only become relevant after the aggregation at the CP.

The estimated amounts of storage that are needed are in the order of 10-100 TB.

ICOS will make use of EGI and PRACE for offering compute services to ICOS data users. The ICOS infrastructure itself does not need any large computation resources.

Figure 8 shows several points in the ICOS technical infrastructure where EUDAT services can play a role.

ICOS Uptake of EUDAT Services

Plans for managing data at the Carbon Portal will be relying on EUDAT services:

B2SAFE: B2SAFE will be used for persistency of the aggregated observational data and elaborated data products at the CP. Current estimates for storage needs of ICOS - produced data are (per year) about 4 TB of data, growing over the coming 3 - 5 years to around 8 TB annually. The elaborated product data volume is more difficult to estimate, but it is likely to start out at 10 TB annually, potentially growing to about 100 TB per year over the coming 3 - 5 years.

B2FIND: Although the main ICOS data discovery services will be offered by the CP, ICOS wants to publish and share the metadata via the B2FIND service and automatic upload or OAI-PMH harvesting of metadata for the observational and elaborated data by B2FIND will be worked on.

B2STAGE: ICOS wants to support end users of its observational data products for instance by facilitating their use as input for computations. The atmospheric scientist group of users are modelling the behaviour of greenhouse gas sinks and sources with models that require large data sets aggregations which also include ICOS data. The potential application of HTC “cloud” computing services (primarily those offered by EGI) to atmospheric modelling will be investigated as part of EUDAT WP7. If that is successful, ICOS will set up a service for selected users whereby ICOS B2SAFE storage and B2STAGE are used to transfer data to and from external computational resources. In addition ICOS will also require HPC resources for larger - scale atmospheric models and therefore collaborations will also be set up with e.g. PRACE or similar networks.

EUDAT DTR: ICOS will be using the experimental EUDAT Data Type Registry (DTR) and add community data type definitions to this registry; this will strengthen interoperability with other disciplines and can help in the creation of an ontology.

ICOS Service Development Requirements

B2SAFE: For (semi-) automatic ingestion of observational data into B2SAFE, ICOS would need to be able to pass user-specific parameters (such as metadata) and return more verbose information (like checksums and PIDs issued by the B2SAFE system). In addition, the development of a “support service” that would simplify the exchange of information about the PIDs and storage locations of the data replicas is needed – this is also necessary in order to provide provenance information about the replicas. Finally ICOS wants there to be support for metadata upload within the B2SAFE service, especially for the ICOS metadata schema.

B2HANDLE: Currently ICOS will mint its own DOIs via the Swedish National Data Service (SND) organization, which is the Swedish DataCite partner. However, ICOS is also interested in the developments for a “unified” B2HANDLE service that can process PIDs from several different PID services, such as EPIC and DataCite (for DOIs).

B2ACCESS: The integration of ICOS-operated services with B2ACCESS (in such a way that e.g. eduGain registered users have access to ICOS RI services) is being considered. ICOS has developed its own AAI solution that functions sufficiently well, so this does not have high priority but ICOS is interested in the further

development of B2ACCESS and may choose to adopt it later - either as a customer of the service operated by the CDI, or by deploying its own local service based on the B2ACCESS code.

B2NOTE: In order to enhance the interoperability of our data and data delivery services, ICOS recognizes the great value of ontology - based approaches to describing the relationships between e.g. data set contents, their provenance and the ICOS organizational entities that produced the data. ICOS has initiated work on creating an ontology, and will also develop e.g. controlled vocabularies to enhance search interfaces.

With respect to semantic services, ICOS would like access to “consulting” services from EUDAT experts in the fields of metadata, semantics and ontologies to support the design and implementation of a registry of all ICOS - specific data types, with the aim of including not only data object types and formats, but also all relevant types of data production techniques (measurements and modelling), and types of (physical) variables contained in the data objects.

8. UPTAKE PLAN OF THE ENES COMMUNITY

ENES Infrastructure Goals and Organization

The European Network for Earth Systems modelling (ENES) provides services for climate and earth systems modelling. The ENES network launched in 2001, and is built on a Memorandum of Understanding (2006) that has been signed by 47 partners (as of 2015) from the academic, public and industrial spheres.

This community is strongly involved in the assessments of the Intergovernmental Panel on Climate Change (IPCC) and provides predictions on which EU mitigation and adaptation policies are based. IS-ENES (the EC-funded infrastructure project of ENES) operates the European part of the global Earth System Grid Federation (ESGF) that offers access to more than 2 PB of climate data.

Many ENES activities are directed towards managing data produced by a series of coupled modelling projects. The World Climate Research Programme (WCRP) working group on coupled modelling promoted a series of Coupled Model Projects (CMIP) whereof the latest phase “CMIP6”, is expected to produce its first experiment data in 2016. The CMIP projects are meant to understand past, present and future climate changes in a multi-model context.

ENES in EUDAT

In the interaction with EUDAT, the following three institutes are involved: 1) Max Planck Institute for Meteorology, 2) Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS), and 3) Deutsches Klimarechenzentrum (DKRZ).

ENES Technical Infrastructure

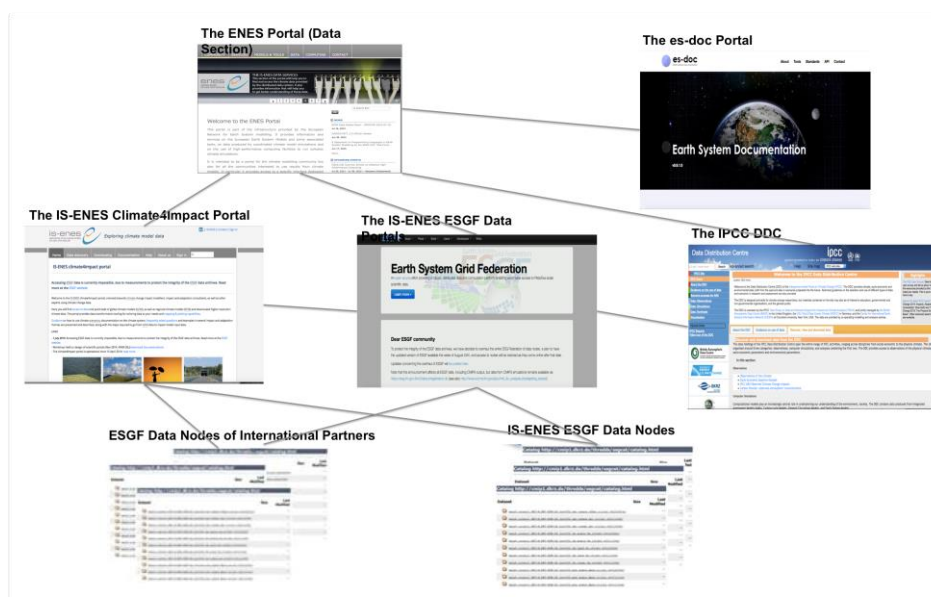


Figure 9: ENES Data Infrastructure showing the relations between the different ESGF and IS-ENES Data Nodes and the different ENES related portals

The ENES technical infrastructure, as part of the global ESGF, is undergoing major changes. With respect to data management, there will be improved versioning and replication management, the introduction of PID services and new processing tools.

The long-term archive at DKRZ (DKRZ-LTA) is the long-term archive of the Intergovernmental Panel on Climate Change Data Distribution Centre (IPCC-DDC) – it stores data from CMIP1, CMIP2, CMIP3 and CMIP5 and the plan is that it will do so for CMIP6 experiment data too. The currently available archiving workflow seems satisfactory for handling that, and no major problems were identified in a gap analysis. The workflows that are in place allow for a thorough check of data objects before they are finally published using a DataCite DOI.

However there is a need for user and project long-tail data archiving.

ENES Infrastructure Requirements

At DKRZ there is already a certified long-term repository in place that meets the need to permanently archive large collections of similar objects from various projects (e.g. CMIP5 and the upcoming CMIP6 with modelling data produced within the context of ENES).

However, users need to be able to preserve, and eventually publish, individual data objects or small collections, either as separate entities or in the context of projects with their own data policies. To meet these requirements, DKRZ would like to deploy and operate a so-called long-tail data repository for its own local users and later on for ENES users. DKRZ would like to maintain its own data publication policy, which requires a tuned quality assessment of data and metadata before a DataCite DOI is assigned to individual objects or collections. This assignment also requires the data to be long-term archived, which in itself establishes certain requirements, like selection of data formats. The quality assessment procedures may vary, depending on the scope and the requirements of projects and users, and need to be defined and implemented in close cooperation between DKRZ (as the maintainer of the repository) and the users/projects.

To limit administrative costs, DKRZ would like to support these – to some extent different – requirements within a single physical repository.

ENES Uptake of EUDAT Services

B2SHARE: The need for user and project long-tail data archiving was mentioned in the previous subsection. DKRZ proposes to deploy a separate instance of EUDAT's B2SHARE service at DKRZ (B2SHARE@DKRZ) to cater for the needs of DKRZ users in the first phase – this would be extended to all ENES users at a later stage.

There are already many specific requirements when it comes to the use of B2SHARE technology by ENES users – a complete list is available in the uptake plan itself. For instance, such archiving requires finely tuned quality assessment before a DataCite DOI is assigned and long-term archiving takes place. Different quality assessment procedures may be needed – these are based on the specific user project requirements, and, for cost reasons, they should all be implemented within a single B2SHARE service. The B2SHARE@DKRZ service will also be targeting data-types that are not well catered for at present, e.g. 'shape files' and additional materials (like figures, images, and grey literature), and that are created by modelling workflows. Finally DKRZ plans to have the B2SHARE@DKRZ repository certified with the Data Seal of Approval (DSA), just as is being investigated for almost all of the EUDAT generic service provider repositories.

B2FIND: Sharing of data is a widely accepted principle within ENES and the climate and the earth system modelling community in general. Research data produced within ENES is not only of interest in the modelling community itself, there is significant interest within multiple other research disciplines. Therefore ENES wants to make such data public and available to other interested parties. These users may use tools and information available from ENES directly, but, based on the experience of ENES to date, it is often advisable to use multiple channels to distribute information and data.

The B2FIND service is an excellent tool for making data widely available since, thanks to its multidisciplinary approach, the service addresses a huge range of research communities. So far metadata from all CMIP projects, including the most recent one (CMIP5), has been made available to B2FIND. ENES intends to make even more metadata available once the relevant data is published. The most prominent candidate for this is CMIP6, but there are also other MIP projects (like OBS4MIPS) that may be of great interest to many researchers.

B2STAGE: From the WP8 discussions, a use case for B2SAFE/B2STAGE came up – it involved staging data to HPC/HTC service sites. This will need to be corroborated and added to the CUP.

ENES Service Development Requirements

B2SHARE: The B2SHARE use case that was described in the earlier subsection will also involve some additional development requirements that have not been analysed properly as yet.

B2HANDLE: With respect to using PIDs, ENES is currently developing and streamlining its part of the global ESGF data infrastructure in view of the needs of the CMIP6 collaborative effort. After evaluating some preliminary demonstrations, in 2015 the ESGF executive committee (ESGF XC) decided to propose federation-wide registration and operational use of Handle-based Persistent Identifiers for all CMIP6 output. The research community responsible for the CMIP6/IPCC experiments also approved the provisional plan. ESGF projects other than CMIP6 are not covered by the current decisions and development, though ENES has expressed a general interest in this as an option for the future.

ENES has evaluated the B2HANDLE library in relation to all the requirements that resulted from the ESGF XC decision. The latest version of the B2HANDLE library already fulfils most of the core requirements satisfactorily. However there are still some open points that have not yet been addressed, e.g. currently the library is not yet general enough and some EUDAT dependencies should be externalized.

ENES is implementing its core infrastructure services and developing tools for end users based on features that the library offers and will tightly embed the library within the ENES component stack. For example, ENES has started using the library in the CMIP6 dataset errata service and for the ESGF PID landing page service. Therefore, ENES strongly encourages EUDAT to offer the B2HANDLE library as a product in the same way that the B2HANDLE service is made available. Doing so will mitigate the risks involved for the ESGF infrastructure and encourage long-term endorsement at the ESGF governance level.

ENES seeks to collaborate with EUDAT to remedy these points and foster the development of the library.

9. UPTAKE PLAN OF THE ELIXIR COMMUNITY

ELIXIR Infrastructure Goals and Organization

ELIXIR is the European infrastructure for biological information, supporting life science research and its translation to medicine, agriculture, bio-industries and society. The goal of ELIXIR is to orchestrate the collection, quality control and archiving of large amounts of biological data produced by life science experiments. Some of these data sets are highly specialized and would previously only have been available to researchers within the country in which they were generated. As of January 2016, ELIXIR consists of 16 member countries and 3 observer countries. ELIXIR appeared on the ESFRI roadmap in 2006 and is one of the three research infrastructures that have been prioritized by the European competitiveness council in May 2014 on the recommendation of the ESFRI. The ELIXIR Consortium Agreement provides the organisation's legal status. ELIXIR is participating in the CORBEL cluster project.



Figure 10: The ELIXIR partners, the EMBL-EBI organization and the ELIXIR Nodes and Hub Structure

ELIXIR in EUDAT

ELIXIR is represented in EUDAT by the European Bioinformatics Institute (EMBL-EBI), which is a major centre in the ELIXIR infrastructure.

ELIXIR Technical Infrastructure

ELIXIR is currently building the ELIXIR Compute Platform, which intends to leverage the expertise and technology of the EGI federation to bring its resources together. At a low level, cloud and cluster resources will be registered in EGI's operations directory and Information service and will be monitored using EGI's Nagios-based system. ELIXIR will provide its own identity management system within which groups and roles can be managed to establish 'Virtual Organisations' that can be authorised to access individual resources.

ELIXIR Infrastructure Requirements

The technical strategy of the ELIXIR Compute Platform is to reuse technology and services where available, or to adapt these technologies and services to meet ELIXIR's needs. Working with e-Infrastructure providers (e.g. EGI, EUDAT, HelixNebula & GEANT) to provide the networked computing and data infrastructure required to support the infrastructure needs of ELIXIR, alongside ELIXIR's own resource providers, is a primary goal of the ELIXIR Compute Platform.

ELIXIR Uptake of EUDAT Services and Service Development Requirements

ELIXIR's uptake of EUDAT services is, for now, primarily related to a data distribution scenario. The data (re-)distribution process can be handled completely with the use of EUDAT services.

The ELIXIR uptake plan that is presented here focuses on the distribution of reference data sets. It is undertaken by EMBL-EBI on behalf of ELIXIR for EMBL-EBI, but could in the future be increasingly undertaken by ELIXIR nodes. Reference data sets can play an important role in broader scientific workflows, yet the existing tools for managing their definition and distribution are immature. This is not unique to ELIXIR. EUDAT will support EMBL-EBI in developing a data set definition and distribution service – this is expected to meet ELIXIR’s requirements and also to be of benefit to other research communities (e.g. earth science).

The Data Set Distribution Service will be offered as a component in the ELIXIR Compute Platform, distributing reference data sets from the ELIXIR Data Platform and from individual researchers responsible for their creation to the facilities where the data sets are to be analysed.

Individual researchers will use the cluster and cloud resources that are integrated into the ELIXIR Compute Platform to install or deploy their analysis pipelines. The selection of a particular resource may be dependent on the local availability of reference data sets and the researcher’s ability to process their own data sets alongside reference data sets using their own defined software environments.

It is not expected that every cloud/cluster in the ELIXIR Compute Platform will have all reference data sets available at all times. However, the Data Set Distribution service could enable researchers to bring their data and analysis pipelines to cloud/clusters where those reference data sets are available. For this to happen, the service (which is being developed with the support of EUDAT) would need to make popular reference data sets readily available on those cloud/cluster resources and also perform rapid updates on those data sets. A priority for the Data Set Distribution Service is to target sites that offer both local and remote researchers the ability to run their own pipelines.

The first priorities are to find the right conditions and environment to make the data transfer work while sidestepping the AAI dependencies. But, as soon as the external AAI infrastructure requirements (ELIXIR, EGI, EUDAT) become clearer, AAI use will be integrated with the Data Set Distribution service.

In parallel with implementing the Data Distribution service at EMBL-EBI, we will also integrate the service with B2SHARE to offer a ready-to-use EUDAT platform for other research communities to test and experiment with.

10. UPTAKE OF EUDAT SERVICES BY THE DATA PILOTS

As the purpose of EUDAT is ultimately to support the widest possible range of European researchers in managing and handling their research data, the project needs to enrich its contacts with European research communities – not only to increase the uptake of its services, but also to bring in new points of view and make sure that EUDAT’s strategies and development efforts do not become overspecialized with respect to the requirements of those research communities, particularly in relation to the communities that are long standing EUDAT partners. Therefore, as a core strategy for EUDAT to broaden its engagement with research communities, EUDAT employs the vehicles of calls for Collaboration and calls for Data Pilots.

The calls for Data Pilots invite research communities and projects to submit proposals for collaborative studies or projects in which EUDAT would provide the storage resources, computing power and expertise to solve a research data management problem. The applicants are expected to match EUDAT’s time commitment in terms of human resources, that is, the number of person months (PMs) they assign to the proposed Pilot should be the same as the number of PMs from EUDAT personnel. It is also important that the proposals address bona fide research use cases. Applicants are encouraged to write proposals that include the use of new technologies (such as Big Data Analytics or Semantics) and that also adhere to Open Access principles and take up output from global organizations (like the RDA¹², CODATA¹³ and W3C¹⁴) which are involved in solving data management questions.

There was an encouragingly warm response of 24 proposals to EUDAT’s first call for Data Pilots. These came from researchers in a variety of disciplines: 7 came from the areas of earth sciences, energy and environment, 6 from the biomedical and life sciences, 6 from the social sciences and humanities, and 5 from physical sciences and engineering. The studies that were proposed had a potential user audience up to 40,000 people and requested a total of up to 4.3 PB of storage resources. The proposals varied with respect to the amount of effort that would be involved and the time span of the collaboration – they ranged from simple storage requests to proposals to develop common services over a period of 18 months. EUDAT certainly faces an interesting organizational challenge when it comes to accommodating so many new tasks, but it also gives us an excellent opportunity to improve the efficiency of our internal organization for handling such a range of diverse engagement tasks.

New requirements for services that have arisen from the Data Pilot proposals will be added to the requirements that have come from the EUDAT core communities and will be used to give EUDAT a better idea about how to prioritise the latter. To make the enabling process with the Data Pilots more efficient, EUDAT will work with the groups that have proposed Data Pilots to create (mini-)uptake plans. These will be used as a reference for the on-going collaboration work in the Pilot and also as input for EUDAT’s Data and Computing Landscape Description Task 4.1.

10.1. Data Pilot Requirements

Since EUDAT is still in the process of scheduling the Data Pilot work and processing the Data Pilot proposals to produce appropriate sets of requirements, this section presents an overview of the requirements that have been distilled at the time of writing. Further details covering all the requirements that come from this round of Pilots will be available at a later stage. Table 2 and Table 3 are provisional requirements analysis tables for the EUDAT core communities and Data Pilot groups respectively. The tables list the EUDAT services and the types of services or modifications to services that have been proposed (grouped by the overall area of data management to which they belong), and indicate whether the developments are actually proposed or are possible use cases for the research communities and groups. These tables are used as supportive evidence to show that requirements from the core communities have wider applicability, and may also be used to decide if requests for completely new services or modifications are of sufficient general interest to

¹² <https://rd-alliance.org>

¹³ <http://www.codata.org>

¹⁴ <http://www.w3.org>

warrant further work. The information in the tables includes requirements from the current call for Data Pilots, and has been taken from the proposal descriptions and interviews with the groups that submitted the proposals.

The EUDAT interactions with the Data Pilot research groups are relatively recent compared to those with the EUDAT core communities, and therefore EUDAT cannot expect those groups to have a deep understanding of the EUDAT services and EUDAT service strategy yet. To remedy this situation, will continue to organize face-to-face meetings with the pilot groups and will also provide specific counselling where needed. As a result we anticipate that further service requirements will surface from those groups, particularly when it comes to discussing the more detailed work plans.

Figure 11 compares the planned service uptake by the EUDAT core research communities (shown in blue) and that of the communities involved in the Data Pilots (shown in red). The numbers are absolute counts that have not been normalized and hence should be interpreted on the basis of there being 7 EUDAT core communities and 21 Data Pilot groups that are included in the analysis.

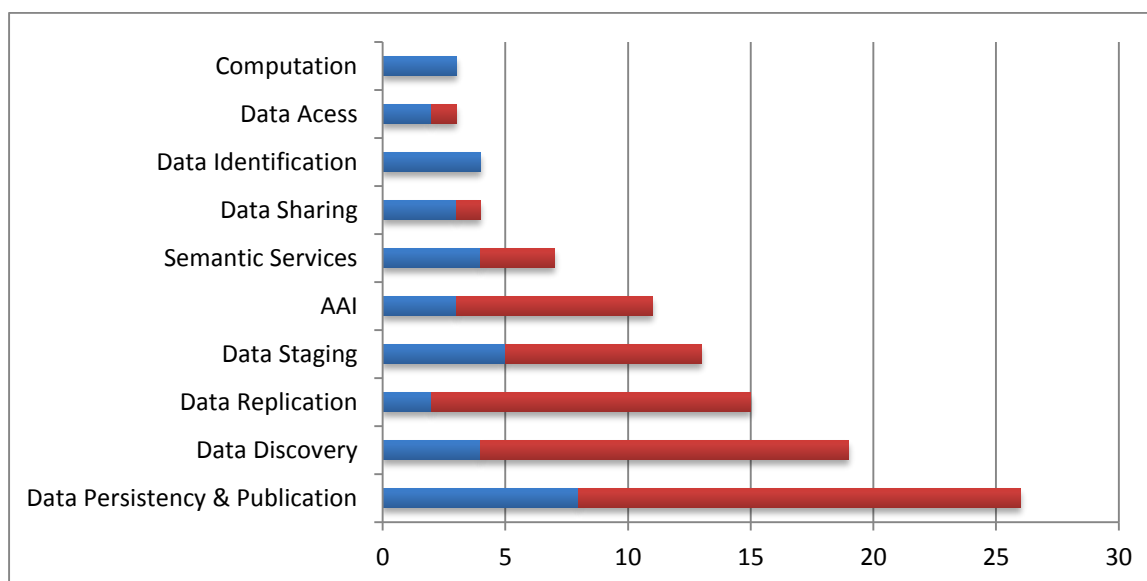


Figure 11: Combined Community (blue) and Data Pilot (red) service selections

From Figure 11 it becomes clear that the Data Pilot proposals focus mostly on data management aspects, such as data persistency and data discovery, using existing EUDAT core services, like B2SAFE, B2SHARE, and B2FIND. Few of the Data Pilot studies proposed using novel services, such as the semantic services. The Data Pilot proposals very rarely go beyond the suggestions for new services in the call, however this was to be expected because we asked the applicants to focus on existing EUDAT services.

It is important to note that almost all of these proposals require only a number of small adaptations to the existing EUDAT infrastructure. However it is also important for EUDAT to guard against specializing the EUDAT services too much. For a description of the strategies that have been developed to meet the Data Pilot requirements in an organizational and resource sense, please see deliverable D6.1.

In summary the current scope of the Data Pilots primarily involves existing data services for replication, sharing, and searching. EUDAT expects that the demand for more sophisticated services will grow as the existing data services are integrated into the infrastructures of the Data Pilot communities. For instance, it is likely that most communities will want integration with the AAI framework later on so they can control access to their data publications.

		CDI ARCHITECTURE PLANS	CLARIN	ELIXIR	ENES	EPOS	ICOS	LTER	VPH
Data Identification (PID)									
	<i>DO part identification¹⁵</i>					X			
	<i>CDI version adm.</i>					X			
	<i>CDI location service</i>	X			X	X	X		
	<i>CDI provenance</i>	X			X	X	X		
Data Persistency & Publication	B2SAFE		X		X	X	X	O	X
	B2SHARE		X		X			X	
	Metadata indexing					X	X		
Data replication	B2SAFE					X			
	<i>B2DISTRIBUTE</i>			X					
Data Access									
	HTTP2replica with Authz.	X				X			
Data Staging	B2STAGE (HPC, EGI)				O	X	X	?	X
Data Synchronization					X				
	B2SAFE&B2DROP Integration					X			
	B2DROP AAI adaptation		X						
AAI									
	Federation integration		X				?		
	B2ACCESS SaaS					X	X		
Data Discovery	B2FIND				X		X	X	
	B2FIND download/staging					X			
	B2FIND as technology or SaaS							X	
Semantic Services	DTR					X	X		
	B2NOTE & B2FIND								
	B2NOTE & B2SHARE							X	
Computation	Deployment of VMs "B2HOST"		X			O		X	

Table 2: An analysis of the (proposed future) EUDAT services used: X (proposed/planned), O (mentioned), ? (possible use case), from the WP8 DLC analysis only

¹⁵ PID part identifiers were suggested useful for modelling dynamic data

		West Life	IST DataRep	Seasonal to Decadal climate and air	Repository for Studnets' results	Enriching EUROPEANA Newspapers	Cloudy Culture	Datapublication@UPORTO	Herbadrop	Unified Access to EISCAT Radar	Tokamak Data Mirror	Data Sphinx	Turbbase	Clinical Trials	Aalto	Clarín user store	NFFA-Europe	Turbulent flows	B2Anno	Global Atmospheric Composition	Fair Data	Ancient OCR
Data Identification (PID)	PIDs																					X
	DO version adm.																					
	DO location service																					
	DO provenance																					
Data Persistency & Publication	B2SAFE	X				X	X	X	X	X	X	?	?					X			X	
	B2SHARE			X	X			X	X	X	X	X	X	P	X		X	X	X	X	X	X
	Metadata indexing	X				X			X	X				X								X
Data Replication	B2SAFE	X																				
	B2DISTRIBUTE																					
Data Access	HTTP2replica with Authz.									X												
Data Staging	B2STAGE	X		P			X		X		X	X	X					X			X	
Data Synchronization	B2DROP	X													X							
AAI	Federation integration				X					X				X	X	X	X	X			X	
	B2ACCESS SaaS													X							X	
Data Discovery	B2FIND	X	X	X		X		X	X	X	X		X	X	X		X	X	X	X		
	B2FIND download/staging																					
	B2FIND as technology or SaaS																					
Semantic Services	DTR												?									X
	B2NOTE & B2FIND																		X			
	B2NOTE & B2SHARE				X			X						X					X			
Computation	Deployment of VMs																				X	

Table 3: A requirements listing of the (proposed future) EUDAT services used: X (proposed) ? (possible use case)

Note that Table 3 includes all the Data Pilot proposals for which EUDAT has sufficient information to be able to flag the use of EUDAT services. Three more proposals are still in the negotiation/clarification stage, so the Data Pilot proposals “City Air”, “SIMCODE” and “Cervical Cancer & Diabetes” are not included in this table because there is not yet sufficient information to classify their requirements appropriately.

11. CONCLUSIONS

EUDAT and the research communities that are currently involved with EUDAT have collaborated on the production of Community Uptake Plans, which give an overview of the plan for the uptake of one or more EUDAT services by the individual communities. These plans can include requests from a research community for EUDAT to provide or develop new or modified services that the community is interested in using. EUDAT is working both with “core” communities that have been partnered with EUDAT over a longer period, and also with research communities that have become involved with EUDAT relatively recently via the EUDAT call for Data Pilot studies.

With respect to the status of the uptake plans for the core communities, a solid basis has been created that allows the EUDAT operations work package, WP6, to start work where the uptake of existing EUDAT services is proposed. All the EUDAT core communities are now in the process of testing and deploying existing services. For further details about this, see deliverable D6.1

When it comes to work on proposals from research communities for EUDAT to provide new or augmented services that were not initially planned, some prioritization of the requests is needed. While some requirements can be taken on board immediately by EUDAT’s service development team from WP5, there are other requests that need further exploration and should therefore be discussed within WP8 and/or the relevant EUDAT Working Groups. EUDAT also needs to take into account that, as well as providing important input to the development of EUDAT services, some of the research communities have expertise in technology development and it can be very beneficial to consider initiating useful development collaborations with those communities as is appropriate.

The new Data Pilot studies offer an important broadening of EUDAT’s scope in relation to the process of gathering requirements from the research communities. The response to the call for Data Pilots has been very positive, and it is also the role of EUDAT’s WP4 (together with enablers from WP6) to ensure that the communication and interaction processes between EUDAT and the communities participating in the Data Pilots are efficient so that EUDAT takes all the community requirements on board.

Integration the EUDAT services more comprehensively with each other in the EUDAT CDI will make the EUDAT services more useful and attractive to research communities. WP5 is currently formalizing the model for data that inherently underlies the EUDAT CDI – they are making it more specific and explicit so it will be more accessible to a wider audience of research communities. This formal CDI data model will also make it easier to discuss the scope of the different services with the various research communities, and simplify discussions with them about the ways in which they connect to the CDI.

For research communities to be genuinely interested in using the EUDAT services in the longer term, it is important that those services will continue to be available and that any costs associated with using the services are clear. Therefore it is vital for EUDAT to have viable models for sustainability and service provisioning. In order to sustain the foundations of the EUDAT CDI beyond the current phase of the project, EUDAT has been working on a CDI Agreement between the service providers that are involved with the project. The EUDAT CDI Agreement formalizes the roles and responsibilities of the service providers constituting the CDI, in particular for those providing EUDAT services. Therefore this Agreement can be seen as a fundamental and necessary step in integrating providers and communities into the EUDAT CDI and will be used to engage with EUDAT’s stakeholders with a view to establishing longer-term agreements and collaboration.

ANNEX A. GLOSSARY

Term	Explanation
AAI	The Infrastructure and services to provide authentication and authorisation
API	Application Programmable Interface
Bitstream	A bitstream is a sequence of bits that encodes a specific informational content, either stored on some media or being transferred under control of protocols
B2ACCESS	Brand of the EUDAT service for federated authentication and authorisation
B2DROP	Brand of the EUDAT trusted cloud storage service
B2FIND	Brand of the EUDAT central metadata catalogue
B2HANDLE	Brand of the EUDAT persistent identifier service
B2HOST	Brand of the EUDAT service to deploy community applications close to the data storage location
B2NOTE	Brand of the EUDAT service to manage semantic annotations
B2SAFE	Brand of the EUDAT service via which the data management policies are implemented within the CDI network
B2SAFE DPM	Brand of the EUDAT B2SAFE data policy manager via which community data manager are able to manage policies which the CDI network from a central portal
B2SHARE	Brand of the EUDAT easy-to-use data repository service
B2STAGE	Brand of the EUDAT comprehensive set of API's and tools to access data managed within the CDI network
B2 service suite	Aggregation name of the EUDAT B2 services
CDI node	Generic of Thematic service provider who has signed the CDI collaboration agreement either as interoperable or integrated partner
CDI	EUDAT Collaborative Data Infrastructure
CDI Gateways	Services or service endpoints (e.g. API or WUI) which are part of the Access Layer of the CDI layered architecture.
CLARIN	CLARIN (Common Language Resources and Technology Infrastructure) ERIC organisation which provides easy and sustainable access for scholars in the humanities and social sciences to digital language data (in written, spoken, video or multimodal form), and advanced tools to discover, explore, exploit, annotate, analyse or combine them, wherever they are located.
Digital Object (DO)	A digital object (DO) is represented by a bitstream, is referenced and identified by a persistent identifier and has properties being characterized by metadata.
DOI	Persistent identifiers within the 10.xxxx prefix namespace domain, issued at a registration agency which is a member of the International DOI Foundation (IDF)
DPM	B2SAFE Data Policy Manager
DSPACE	Is an open source repository software package typically used for creating open access repositories for scholarly and/or published digital content
eduGAIN	The eduGAIN service interconnects identity federations around the world, simplifying access to content, services and resources for the global research and

	education community. eduGAIN enables the trustworthy exchange of information related to identity, authentication and authorisation (AAI).
ELIXIR	ELIXIR is the European infrastructure for biological information, supporting life science research and its translation to medicine, agriculture, bio-industries and society.
ENES	European Network for Earth System modelling developing a common climate and Earth system modelling distributed research infrastructure in Europe
ePIC	Consortium of European partners in order to provide PID services for the European Research Community, based on the handle system (TM, http://www.handle.net/), for the allocation and resolution of persistent identifiers
EPOS	The <i>European Plate Observing System</i> is the integrated solid Earth Sciences research infrastructure.
ERIC	European Research Infrastructure Consortium
ESFRI	European Strategy Forum for Research Infrastructures
ESFRI Roadmap	A regularly published report by ESFRI on the status of European research infrastructures
GridFTP	Is a high-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks
Handle	Technology for the creation of, and access to resolvable unique identifiers, is developed and maintained by CNRI
HTTP	Is an application protocol for distributed, collaborative, hypermedia information systems
ICOS	ICOS, the Integrated Carbon Observation System is a pan-European Research Infrastructure for quantifying and understanding the greenhouse gas balance of the European continent and adjacent regions.
IdP	Organizational Identity Provider within an identify federation
iRODS	Technology for flexible policy based data management of files and metadata that span storage devices and locations, developed and maintained by the iRODS consortium
LTER	LTER, the Long Term Ecological Research Network is a network comprising research sites and platforms to conduct research on ecological issues and better understand ecosystems.
Metadata	Metadata contains descriptive, contextual and provenance assertions about the properties
OAI-PMH	The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) which is a low-barrier mechanism for repository interoperability.
OpenID	Is an open standard and decentralized authentication protocol
ownCloud	Enterprise file sharing solution for online collaboration and storage, developed by the ownCloud company

Persistent Identifier (PID)	A persistent identifier is a long-lasting ID represented by a string that uniquely points to a DO and that is intended to be persistently resolvable to access meaningful, current state information about the identified DO
RDA	The Research Data Alliance (RDA) is a community-driven organization from in the European Commission, the United States Government's National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation with the goal of building the social and technical infrastructure to enable open sharing of data.
RDA-DFT	The RDA Data Foundation and Terminology working group
RDF	Resource Description Framework (RDF) is a family of the World Wide Web Consortium specification originally designed as a metadata model, commonly used as a general method for conceptual descriptions or modelling of information
RESTful	Representational State Transfer (REST) is a software architectural style of the World Wide Web. Systems or services that conform to the constraints of REST can be called RESTful.
SAML	Security Assertion Markup Language is an XML-based, open-standard data format for exchanging authentication and authorisation data between parties, in particular, between an identity provider and a service provider.
Service Provider	Organisation or federation or part of an organisation or federation that manages and delivers a service or services to customers
SRU	Is a standard XML-based protocol for search queries, utilizing CQL – Contextual Query Language
Technical Committee	Governance body to organize the collaboration, knowledge transfers and information flow between the technical-oriented work packages (e.g. Community requirements and engagement (WP4), Service building (WP5), Operations (WP6), Cross e-infrastructure services (WP7), Data life cycle across communities (WP8), Technology exploration (WP9))
Unity IDM	Technology for identity, federation and inter-federation management, is used within the B2ACCESS service
URL	Uniform Resource Locator
VPH	VPH, the Virtual Physical Human is the community that aims at developing integrative biomedicine
Webdav	Web Distributed Authoring and Versioning is an extension of the HTTP that allows clients to perform remote web content authoring operations
W3C	World Wide Web Consortium
XML	Extensible Markup Language is a markup language that defines a set of rules for encoding documents in a format which is both human-readable and machine-readable
X.509	is an <u>ITU-T</u> standard for a public key infrastructure (PKI) and Privilege Management Infrastructure (PMI)