CS3MESH4EOSC

Data Management Plan

Project acronym: CS3MESH4EOSC

Deliverable D1.3: Data Management Plan

Contractual delivery date	31-03-2020
Actual delivery date	29-05-2020
Grant Agreement no.	863353
Work Package	WP1
Nature of Deliverable	R (Report)
Dissemination Level	PU (Public)
Lead Partner	CERN
Document ID	CS3MESH4EOSC-20-004
Authors	Pedro Ferreira (CERN), Jakub Moscicki (CERN), Anna Manou (CERN)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863353

Document Revision History

Version	Date	Description	Author(s)
v1	29-05-2020	Initial version of the DMP	Pedro Ferreira, Jakub Moscicki, Anna Manou
v2	07-12-2020	Added Google Analytics	Pedro Ferreira

Abstract

The deliverable D1.3 Data Management Plan aims at presenting a plan for the Project data and Platform data management. It presents the data privacy practices adopted by the Project for the Project and the actions put in place to ensure a user privacy-friendly platform. The current document is the first version of the Data Management Plan, which will be updated in M34.

Table of Contents

1 Introduction

This document's main goal is to establish a data management framework which will apply to two main types of data:

- **Project Data** - data which is produced during the project itself, usually by members of the consortium, its governing bodies or outsourced third parties. This is data which is needed for the correct functioning of the project as well as actual deliverables resulting from the Project's tasks. Some examples of such data are:
    - Meeting minutes and agendas;
    - Contacts in a newsletter;
    - Survey responses;
    - Personal data stored in cloud services/tools.
- **Platform Data** - data which is produced through the use cases of the project, meaning any systems created in the context of CS3MESH4EOSC and every possible service based on those systems. Although this data, in practice, will most probably not fall under the Project's jurisdiction, it is the latter's responsibility to ensure that it is produced under specific circumstances, respecting a set of standards and as to provide guarantees to any prospective users with regards to their rights and their ability to use the data to their advantage. Examples of such data are:
    - Files stored in mesh nodes or shared using OCM;
    - File metadata synchronized through the CS3 APIs;
    - User data stored in the various ScienceMesh nodes.

In the first part of this document, we will be looking into the data privacy practices which will be put in place within the Project as well as the roles and responsibilities involved. We will analyse the collaborative tools which will be used, as to identify the types of data which they are likely to collect and process, and clarify their privacy practices.

The second part of the plan will focus on making sure that the data privacy rights of the future users of the ScienceMesh distributed platform, the main outcome of the Project, are taken into account in the platform's design. To this end, we shall provide the basis for what should be a framework for the development of a user-centric project with an emphasis on data privacy. We shall also study some of the fundamental use cases of the Project. This last part will be gradually expanded as the project's requirements solidify and new considerations arise.

The Data Management Plan is a "living document", which means that it will keep on being updated throughout the Project's lifetime.

2 Project Data

As explained in the introduction of this document, the concept of "Project Data" refers to all data which is produced in the context of CS3MESH4EOSC, by members and bodies of the consortium or third parties who may be commissioned by the Project to execute activities in its context. This data may include Personally-identifiable information (PII) covered, in EU and EEA territory, by the General Data Protection Regulation (GDPR) and equivalent legislation elsewhere. Some examples of such data might be:

- Meeting minutes and agendas;
- Internal Project sites with lists of persons, contacts and other PII;
- Project data stored in cloud services/tools;
- Contacts of members of a newsletter;
- Transcripts of interviews done to partners and power users.

In order to certify that the rights of every data subject are respected, it is essential to make sure that the Project has the appropriate control mechanisms and best practices in place, as well as a clear definition of the roles which the various Project bodies will play.

2.1 Roles and Responsibilities

The management and protection of Project-related personal data is, first of all, a responsibility of every single partner in the CS3MESH4EOSC consortium. All involved organizations should make sure that their practices, tools and processes are compliant with their local and organisational data privacy regulations and proceed to any clarifications with their own Data Protection Officer (DPO) or equivalent body, if needed. This is stipulated by point 11.6.1 of the Consortium Agreement:

> *"All personal data processed by the Parties for the purpose of the Project shall be processed in accordance with their respective legal frameworks".*

Within the CS3MESH4EOSC Project there shall be, however, control mechanisms to ensure that any possible risks relating to this matter are properly monitored and addressed, and that the information contained within this document remains up-to-date:

- **Work Package Leaders** shall be responsible for ensuring that any tools used in the context of their WPs have reasonable privacy policies which guarantee the privacy rights of every individual working in them, but also of any other parties covered by the usage of such tools in the context of CS3MESH4EOSC. Some examples of such activities may be: management of mailing list contacts or newsletter membership, recording of interviews for outreach or even the collection of data through surveys. Any newly adopted tools should be the object of careful evaluation of its policies and the StC should be notified accordingly;
- The **Steering Committee** shall be the main communication channel used to discuss such tools and otherwise clarify any questions regarding Project-related data. Whenever doubts of a legal/organisational nature persist, they should be relayed to the DPOs of the concerned organizations;
- The **Project Coordinator** shall be responsible for ensuring that the overall data privacy practices and guarantees are adequate and match the level stipulated by CERN's internal legislation[1];

[1] Point 11.6.1 of the Consortium Agreement: *Such processing of personal data shall ensure at least the level of data privacy stipulated by the Coordinator's internal legislation.*

- The **Work Package 1 Leader** shall be responsible for ensuring that the DMP remains up to date.

2.2 Protection of Personal Data

Partners are allowed to process any personal data only throughout the lifetime of the project. After the end of the project, partners can retain personal data for a period of five years, whenever it is needed for justification purposes towards the European Commission. At the end of the five-year retention period, partners are obliged to permanently delete all personal data in their possession.

In case of data breach, the data controllers should report it to the Steering Committee and the Project Coordinator. The liability for the data breach falls on the Data Controller, who is also responsible for reporting it to the relevant supervisory.

It is important to highlight that no partner is allowed to share any personal data with third parties unless they obtain the consent of the remaining members of the Consortium or the express consent of the data subjects. As the Consortium Agreement states in section 11.6.2:

> *"Personal data shall not be further transferred or disclosed to third parties unless otherwise agreed by the Parties in writing, or unless consent of the data subjects has been obtained in advance".*

2.3 IT Services

The establishment of internal communication and collaboration tools falls within the scope of Work Package 1, as per task T1.3 ("Internal communication and organisation of meetings"), which is responsibility of CERN. Due to this fact, three out of the six collaborative tools officially adopted by the project are actually provided directly by this organization, two of them being already existing services which are known to be reliable and widely used both inside and outside CERN (Indico and CERNBox). The Zoom Video-conferencing System used by the project is a cloud service provided through DTU, which possesses a license for its own activities.

CERN's own Data Privacy Policy is defined by the organization's "Operational Circular No. 11" (OC11)2, which establishes the general principles for the collection and processing of personal data at CERN as well as the obligations of the organization with regards to the "data subjects".

As for the Zoom service used by DTU/DeiC to set up the project's video-conference meetings, it claims compliance with the GDPR, through a Data Processing Addendum (DPA)[3] to its Terms of Service, which guarantees to all its users the rights granted by the GDPR. Zoom Video Communications, Inc. also takes part in the EU - U.S. and SWISS - U.S. Privacy Shield Framework[4], which guarantees compliance with data protection requirements in data transfers "across the Atlantic". The European Commission deemed the EU - U.S. Privacy Shield Framework adequate to enable data transfers under EU law[5]. The same applies to GitHub, which claims compliance through the aforementioned framework.

[2] https://cds.cern.ch/record/2651311
[3] https://zoom.us/data-processing
[4] https://www.privacyshield.gov/participant?id=a2zt0000000TNkCAAW&status=Active
[5] https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/eu-us-data-transfers_en

2.3.1 Privacy Policies

2.3.1.1 Indico Meetings

Controller	CERN	**Processor**	CERN
Applicable Policies	CERN's OC11	**Transfers**	None
Collected Data	• Meeting participant names/e-mails • Participant presence/absence information • IP addresses of users (logging) • User profile information (title, name, e-mail, affiliation, phone number and address) • User preferences and settings		
Retention Periods	• Participant information - **until deleted by organizer** (kept for data preservation purposes, restricted to consortium) • User profile information - **lifetime of user account** • IP addresses (logs) - **13 months**		

2.3.1.2 Wiki

Controller	CERN	**Processor**	CERN
Applicable Policies	CERN's OC11	**Transfers**	None
Collected Data	• Personal data (name, e-mail and username) • User settings • GitHub account information (username/e-mail)		
Retention Periods	**For the duration of the project** (only text data and files will be preserved afterwards)		

2.3.1.3 Zoom Video-conferencing System

Controller	DTU	**Processor**	Zoom Video Communications, Inc.
Applicable Policies	EU - U.S. Privacy Shield	**Transfers**	As defined by Zoom's Privacy Policy[6]
Collected Data	From Zoom's Privacy Policy: • *Information commonly used to identify you, such as your name, username, physical address, email address, phone numbers, and other similar identifiers* • *Information about your job, such as your title and employer* • *Credit/debit card or other payment information* • *Facebook profile information (when you use Facebook to log-in to our Products or to create an account for our Products)* • *General information about your product and service preferences* • *Information about your device, network, and internet connection, such as your IP address(es), MAC address, other device ID (UDID), device type, operating system type and version, and client version* • *Information about your usage of or other interaction with our Products ("Usage Information")*		

[6] https://zoom.us/privacy

	• *Other information you upload, provide, or create while using the service ("Customer Content"), as further detailed in the "Customer Content" section below*
Retention Periods	Not clearly specified. From Zoom's Privacy Policy: *We will retain your Personal Information for the period necessary to fulfill the purposes outlined in this Privacy Notice unless a longer retention period is required by law.*

2.3.1.4 CERNBox

Controller	CERN	**Processor**	Consortium
Applicable Policies	CERN's OC11	**Transfers**	None
Collected Data	• User information (name, e-mail, username) • User settings • IP addresses (logs)		
Retention Periods	• User information - **indefinite** • IP addresses - **13 months**		

2.3.1.5 GitHub

Controller	CERN	**Processor**	GitHub, Inc.
Applicable Policies	EU - U.S. Privacy Shield	**Transfers**	As defined by GitHub's Privacy Policy[7]
Collected Data	From GitHub 's Privacy Policy ("short version")[8]: *GitHub collects information directly from you for your registration, payment, transactions, and user profile. We also automatically collect from you your usage information, cookies and similar technologies, and device information, subject, where necessary, to your consent. GitHub may also collect User Personal Information from third parties. We only collect the minimum amount of personal information necessary from you, unless you choose to provide more.*		
Retention Periods	• For **as long as the account is active**[9]. • Code authorship information: **indefinite**		

2.3.1.6 Gitter

Controller	CERN	**Processor**	GitLab, Inc.
Applicable Policies	GDPR	**Transfers**	Google Analytics, Mandrill/Mailchimp, Transloadit
Collected Data	• User profile information • Chat messages		
Retention Periods	• Messages remain **as long as the author/room administrator doesn't delete them**		

[7] https://help.github.com/en/github/site-policy/github-privacy-statement#how-we-share-the-information-we-collect

[8] Full version: https://help.github.com/en/github/site-policy/github-privacy-statement#what-information-github-collects

[9] https://help.github.com/en/github/site-policy/github-privacy-statement#data-retention-and-deletion-of-data

	• User account can be "ghosted" (dissociated of personal information)

2.3.1.7 Google Analytics

Google Analytics is used in the CS3MESH4EOSC official Project website, in order to measure the impact of this particular communication channel.

Controller	Trust-IT	**Processor**	Google Inc.
Applicable Policies	GDPR	**Transfers**	Google (US) through Contract Clauses[10]

Collected Data	• User-level data[11]:
	  ○ User's unique identifier (cookie);
	  ○ Online identifiers, including cookie identifiers;
	  ○ IP addresses;
	  ○ Device identifiers;
	  ○ Client identifiers;
	• Event-level data (on-site/app activities, such as "page views" and "user actions")[12].		

Retention Periods	• User-level data – 14 months; • Event-level data – 38 months;		

2.3.2 Other Services

While the Project recommends a set of services which are known to work and, as demonstrated above, offer reasonable guarantees with regards to Data Privacy, this not by any means an exhaustive list of every possible such service which shall be used in the Project's lifetime. As different needs arise, other tools may have to be employed, which should be evaluated with regards to the aforementioned parameters. Being a "living document", the DMP should be updated accordingly.

As mentioned above, Work Package leaders are responsible for making sure that whatever tools are used in their WPs have been properly evaluated and constitute no threat to the privacy rights of all data subjects involved. The StC should be promptly informed about any newly adopted such tools, especially if they are intended to collect and/or process data of persons who are not part of the Consortium.

[10] https://privacy.google.com/businesses/processorterms/mccs/
[11] https://developers.google.com/analytics/devguides/collection/analyticsjs/cookies-user-id
[12] https://support.google.com/analytics/answer/1033068

3 Platform Data

The CS3MESH4EOSC project aims to provide the world with the base technology which will allow for the creation of interconnected meshes of EFSS systems. The project's goal is not mainly to define who will be part of such a data ecosystem (although we believe we have a critical mass which will be essential in bootstrapping the community) but rather to define how such an ecosystem will work - what will be the basic workflows, what will be required from mesh nodes and how end users will interact with them. These workflows are obviously inseparable from the technology which underlies the whole project: CS3APIs, OCM, REVA and others. After all, they will limit what end user applications will be able (or not) to do with the various types of data which are collected and stored.

Although the data which will be generated, collected and stored in the context of future services based on the results of this project is outside of our control and out of the scope of the Project, we thought it would be important to establish what the expectations for a service of this kind would be. After all, as already mentioned above, the technology output of CS3MESH4EOSC (ScienceMesh) will greatly condition the functioning of any platforms potentially based on it. It is thus essential to clarify the principles behind ScienceMesh and the way they shape the flow of data between nodes.

One aspect which cannot be possibly ignored in a system of this nature is that of data privacy. Never in the history of ICT has there been such widespread concern regarding the safety and exploitation of personal data. From the political and legal standpoint, the GDPR has set an important milestone which represents a shift in what is expected from service providers in this particular matter - the latter are bound to transparency regarding their operations and are now *de jure* responsible for the proper handling of the data which users entrust to them.

As a "child of its time", ScienceMesh intends to not only abide by current data privacy regulations, but to incorporate the principles as well which are subjacent to them in its own DNA. By treating data privacy as first-class citizen, we hope we will provide a solution which some would describe as "squaring the circle": it respects everyone's rights while remaining useful, convenient and hassle-free for its users and providing clear value to any services based on it. We expect we will provide a tool which will allow organizations to share and collaborate without sacrificing any of their data sovereignty - not only a sovereignty based on trust and on the letter of the law, but also a practical sovereignty enforce through privacy-preserving technology.

3.1 Privacy by design

Privacy by design[13] is an approach to system design which calls for privacy to be taken into account throughout the whole project. Since the adoption of the GDPR in 2016, it became the recommended default choice in the design of information systems, so that *"by default, only personal data which are necessary for each specific purpose of the processing are processed"*[14].

From a broader point of view[15], the *Privacy by Design* approach states seven "foundational principles":

- P**roactive, not Reactive; Preventative, not Remedial** - anticipating privacy-invasive events before they happen;
- **Privacy as the default setting** - if a user does nothing, their privacy remains intact;

[13] Hes, R. & Borking, John. (1995). *Privacy-Enhancing Technologies: The Path to Anonymity.*
[14] https://gdpr.eu/article-25-data-protection-by-design/
[15] Cavoukian, A. (2011) *Privacy by Design - The 7 Foundational Principles - https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf*

- **Privacy embedded into design** - privacy as part of the core of the application;
- **Full functionality - positive-sum, not zero-sum** - it should be possible to have the best of both worlds (e.g. privacy vs. security);
- **End-to-end security - full life cycle protection** - "cradle to grave" secure life cycle management of information (e.g. automatically disposed of at the end);
- **Visibility and transparency - Keep it open** - subject to independent verification;
- **Respect for user privacy - keep it user-centric** - the user is the most important part of the system.

As already mentioned above, given the early state of some of the technologies developed in the context of the CS3MESH4EOSC project and the current social context with regards to the storage and processing of personal data, the only logical choice was to adhere to these principles and to integrated them into the foundation of the project.

We are also conscious that in certain cases, a balance will have to be struck as to not jeopardize too much the user-friendliness of the system. In such cases we will still try to fall back to a "privacy as the default setting" scenario, with less optimal (but nonetheless useful) features added as "opt-in" modules. That will not an easy task by any means, but it is one on which we believe the future viability of the project depends greatly.

4 ScienceMesh Use Cases

In this section we will go over some of the use cases which have already been defined in the context of the Project and see how they affect the pathways data takes within the mesh and which consequences they have in data storage and user privacy in particular (with a privacy-by-design approach in mind).

While the "Roadmap for Trust" task (T2.1) is still being worked on at the time of writing of this document, we have based the descriptions below on a working version of the document[16]. There is, of course, the possibility that they will change and we will take that into account in future revisions of this plan.

4.1 Sharing files across organizations

This use case is the arguably the most fundamental workflow, which demonstrates the principle behind ScienceMesh: two users, Albert and Peter, are on distinct cloud nodes run by their corresponding institutes (CERN and CESNET, respectively). Peter wants to share some files with Albert.

According to this workflow:

1. CESNET's server will contact and negotiate a shared resource with CERNBox (CERN's ownCloud instance)
2. Albert will then be notified by CERNBox of Peter's intention
3. Albert will be able to to access the resource in question

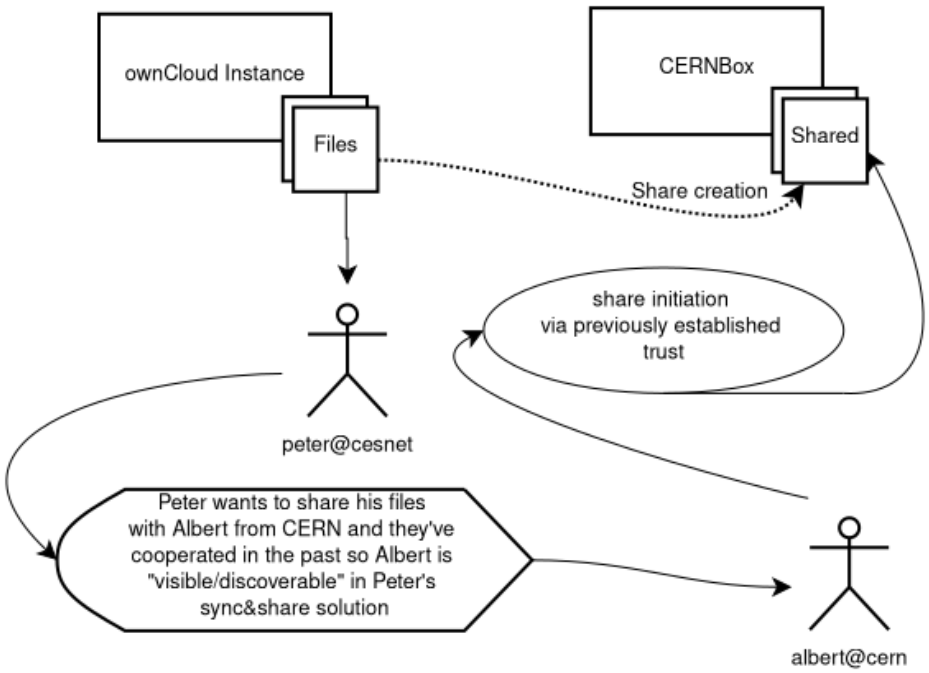

Figure 1: Sharing files across organisations workflow

[16] https://wiki.cs3mesh4eosc.eu/wp2/wp2_-_task_2.1.pdf

This workflow takes for granted that the source system (CESNET) already knows how to refer to the target user on the target system. This can be achieved through some sort of unique identifier, which can be either local or global. This is the only piece of information which has to be known by the source system about the user in question. Getting this identifier is something which will be discussed in the next two use cases.

From	To	Content	Sensitivity
Source system	Target System	Invitation to share data (target user's unique ID, name of share)	Medium

Table 1: Sharing files across organisations

4.2 Initiation through side channel

The second use case we will illustrate involves sharing files with a target user whose unique identifier is unknown to us. Taking the same actors as the above example into account, Peter will have to somehow find out which unique identifier is held by Albert. The naive approach to this problem would be the typical "phonebook" solution, which we will actually hint at in the next use case. However, we do want to be consistent with the **privacy as the default setting** mantra which was mentioned before, which implies the existence of a solution which can allow the two parties to negotiate the exchange of the unique identifier through a channel which they consider safe. The easiest approach is for Peter's host system to generate an anonymous "invitation link" which he can then send to Albert for instance via e-mail. The process would be the following:

1. Peter generates an anonymous invitation link using his own system (CESNET ownCloud). This link contains a unique random token or alternatively a digital signature;
2. Peter e-mails the link to Albert;
3. Albert receives the link. At this point, we will assume he will go to his home system (CERNBox) and paste the link in some sort of "add external share" dialog (see below for another option);
4. CERNBox takes the link and extracts the token, which it can send to CESNET ownCloud to prove it holds the invitation from Peter;
5. CERNBox confirms that the token/signature is valid and establishes the share with CESNET ownCloud;
6. Both parties get each other's unique identifier as part of the process.

In this way, Albert's host system will never know anything about Peter other than his unique identifier (and vice-versa). As a matter of fact, the unique token mentioned above (or another one negotiated based on it) is all which is needed to allow the two parts to establish trust on shared content. However, we do want to cater to the initial use case, by allowing Peter and/or Albert to establish future shares with no need to undergo the link-based sharing procedure once again.

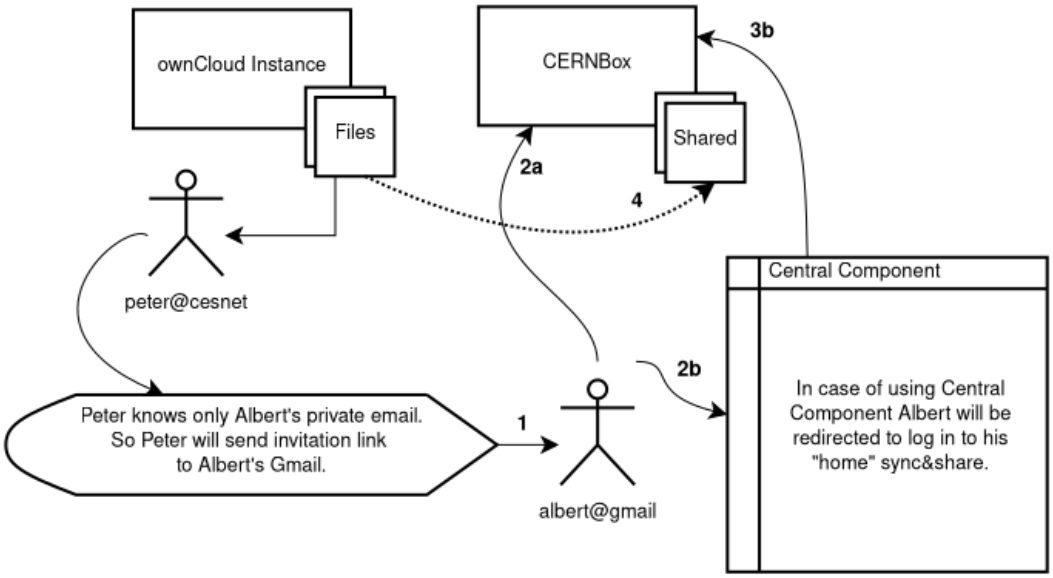

Figure 2: Initiation through side channel workflow

From	To	Content	Sensitivity
Source user	Target user	Link containing one-time “invitation” token	Low
Target system	Source system	Invitation token	Low
Source system	Target system	Source user’s unique identity	Medium
Target system	Source system	Target user’s unique identity	Medium

Table 2: Initiation through side channel

4.3 User Directory as an opt-in feature

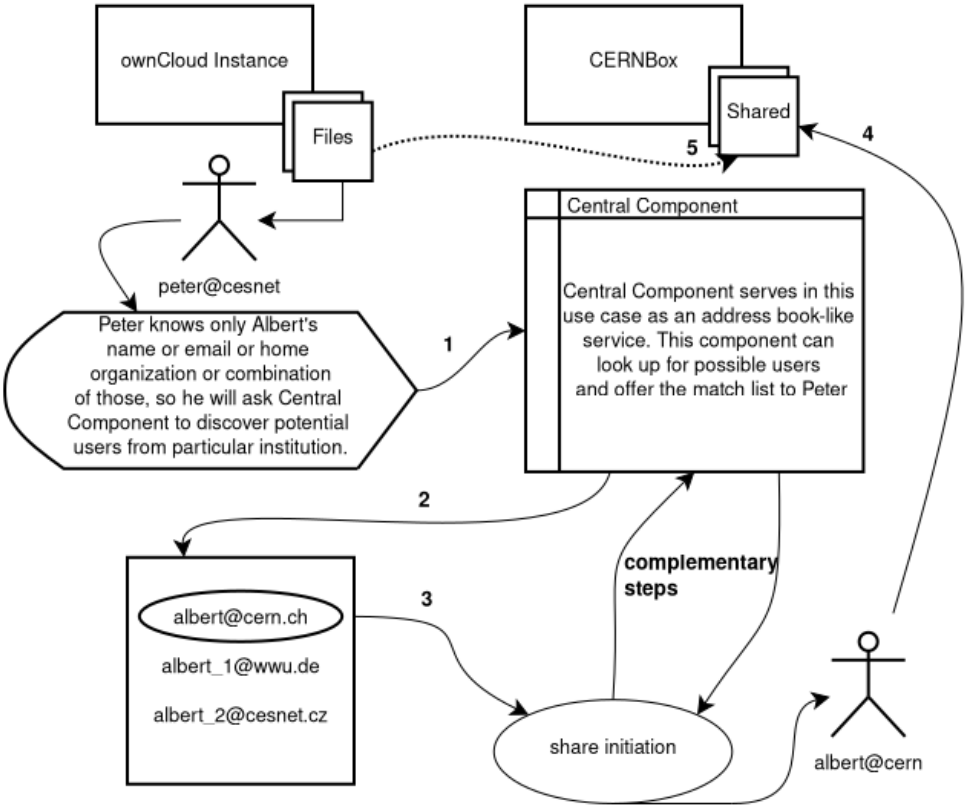

Figure 3: User Directory as an opt-in feature workflow

The last use case we will be documenting builds on the aforementioned metaphor of the "phonebook". This is a scenario which will require the existence of an additional component which will be external to the two participating mesh nodes. This "component" will act as a user directory, effectively allowing user identifiers to be discovered from other pieces of information (e.g. user's name, e-mail, or even organization).

The workflow would be the following:

1. The first time that Peter logs in to his ScienceMesh-based service at CESNET, he is asked whether he would like to opt in to be listed on ScienceMesh's global user catalogue;
2. Albert looks for Peter, using the "mesh search" function with CESNET's ownCloud service provides. He knows Peter's name, so he manages to identify him;
3. The "mesh search" function obtains Peter's unique identifier from the user catalogue;
4. *The rest of the sequence is equivalent to the "sharing files across organizations" use case.*

It is worth mentioning once again that, in order to be consistent with the principle of **privacy as the default setting**, being listed in such a "user catalogue" would be a decision which would have to be explicitly taken by the user ("opt-in" rather than "opt-out").

From	To	Content	Sensitivity
Target user	Catalogue service	User information: name, organization, email (?)	High (personal data)
Source user	Catalogue service	Search parameters	High
Catalogue service	Source system	Target user's unique identity	Medium
Source system	Target system	Invitation to share data (target user's unique ID, name of share)	Medium

Table 3: User Directory as opt-in feature

It is clear from this summary of the data flow that the central "catalogue service" which is introduced above would end up aggregating a considerable amount of personal data. This means that special care should be taken with it, as to make sure that only the bare minimum amount of personal data is stored. It is also essential that the amount of data which is displayed back in the form search results is a reduced subset of what is stored. For instance, while searching by e-mail may be useful, the search results should never include the e-mails of users.

The existence of such central infrastructure will also require the establishment of policies and guidelines which both participating sites and the Project's infrastructure itself will have to adhere to. In this respect, the outcomes of Tasks 2.2 (Operational excellence and new sites) and 2.5 (Security in federation) will be essential in laying the foundation for ScienceMesh's governance throughout the later phases of the Project and after it is over.

4.4 Use Cases - Conclusions

We have described the three basic use cases which will be at the heart of collaboration in what we hope to be a pan-European (and potentially worldwide) privacy-preserving ScienceMesh. We believe our approach to them will manage to be consistent with the Privacy by Design principles we have committed to. A quick run-through of the principles mentioned above shows that:

- We are trying to be **proactive**, studying the use cases carefully and trying to foresee their implications;
- We are aiming for **privacy as the default setting**, as the use cases above demonstrate - loss of privacy will only happen if users opt into it;
- Privacy is effectively **embedded into the design** of the system, all exchanged information is carefully tracked and approaches involving anonymous negotiation are favoured;
- We are aiming for **full functionality**, with all features of the system available by default to users even in the most restrictive scenario;
- **End-to-end** "cradle to grave" secure management of information is being incorporated in the Project's design and development processes, with the WP2 leader as the person responsible for liaising with other packages, with the advice of the PC and Technical Coordinator;
- **Visibility and transparency** are not an issue in this Project, which is Open Source from its inception and whose documentation and deliverables will be public;
- We are committed to making the system as **user-centric** as possible, which is why the project puts an emphasis on establishing communication with prospective user communities as well as current users of the services taking part in the Consortium.

5 Index of Tables

6 Table of Figures