EPOS Competence Center USE CASE

# 1. Introduction

The purpose of this document is to outline the pilot use case be developed within the framework of EGI EPOS Competence Center

# 2. Main topic

The main topic of this use case is the implementation of AAI modules within already existing e-Infrastucture, **both for discovery and processing purposes**.

# 3. Background

The EPOS community is organized in

1. National Layer, made up of Research Infrastructures (RIs) providing data and services,
2. Community Layer, made up of pan European e-Infrastructures which disseminate data and services of a single discipline (e.g. seismology with ORFEUS/EIDA),
3. Integration Layer, (Integrated Core Services - ICS) the e-Infrastructure designed and ran by EPOS; this is the place where the integration of data and services provided by the Thematic Core Services (TCS, Community Layer) occurs.

In this complex framework the major challenge is to enable scientists to make use of multidisciplinary data, which are usually heterogeneous in terms of formats, metadata and accessibility.

Integrated Core Services are characterized by a **Central Hub (ICS-C),** whose main goal is to host the metadata catalog and orchestrates external resources (e.g. HPC), and the D**istributed Services (ICS-D),**  whose goal is to provide resources (e.g. computational, visualizaiton).

**For the development of this use case we will focus on the EPOS ICS-C layer, its interaction with the EPOS TCS layer and its interaction with the distributed computational services (ICS-D).**

# 4. The use case

The present document takes into account a basic use case, which deals with:

1. **A basic multidisciplinary use case**, dealing with the discovery of heterogeneous data from the ICS-C portal by a user who wants to discover and access (e.g. download) such data. This use case represents the minimal objective in the framework of the present EPOS CC,
2. **An extended, single discipline, computational oriented use case**, dealing with the usage, from user’s side, of a computational seismology tool which orchestrates the access to data and to computational resources on behalf of the user

Scenario and actors of both use cases are reported in the following diagram.



## 4.1 Basic use case: discovery, access and delegation

### Description.

In this simple use case, the user access to the ICS portal, discover and selects some data, which is downloaded on his/her computer.

### Prerequisites

1. ICS is registered as a service provider in the federation of the used AAI mechanism. Two different AAI mechanisms can be used, one for the interaction users/ICS and the other for the interaction ICS/TCSs.
2. The user has to be represented with the same common Id in the 2 AAI mechanisms.
3. The user is granted with the lowest level of assurance
4. Metadata to perform discovery is ingested and queryable at the EPOS metadata catalogue level.
5. TCSs have web services implemented and some kind of AAI system, which “knows the user” (e.g., they all belong to the same realm/domain or have common Id provider).
6. Users can register on ICS only. The ICS provide users with the unique common Id after the completion of the registration process. ICS should grant users with authorisation to access TCS services.
7. ICS could provide an IdP for homeless users.
8. There is an AA(A)I on ICS. Accounting is in parenthesis since it is not necessarily right now.
9. TCSs constantly update metadata catalogue on ICS
10. No data, except metadata resides on ICS
11. TCSs can understand “identities” from ICS, together with attributes concerning data permissions, access, etc.

### User workflow

1. In case this is the first time the user access to ICS-C portal, s/he registers to EPOS ICS-C user database.
2. User does some simple but multidisciplinary data discovery (i.e. accessing to at least two types of data from different domain and TCSs, say seismological waveforms from seismology TCS and events from AH TCS)
3. S/he gets the complete list of results (e.g. data-objects, files in this case) and selects some 2 of them to be downloaded
4. S/he obtains the data (e.g. download as zipped/tar format or simply in the native file format).

### System workflow

* 1. Once the user is authenticated, the IdP redirects user to ICS-C with a TOKEN
	2. ICS-C assigns additional attributes to perform default actions
1. 1. ICS-C performs discovery and receives the selection of 2 or mode files, stored in different TCSs, from the user
2. 1. ICS contacts TCS1 and TCS2
	2. TCSs check user’s permissions to access the files through the attributes assigned to the users
	3. After permission check, TCSs send a temporary URLs to download the data to the ICSs
	4. ICS shows the temporary URLs to the user
	5. User can start the downloads through the given URLs (direct connection with TCSs)

### Comments, and proposed technical overview:

1. **Authentication**

It is assumed that this step is done by any (but reasonable) authentication service including eduGAIN, OpenID and (yes!) Google, Facebook and others if needed.

1. **Affiliation check** (not important for above scenario but crucial for EPOS)

Before we proceed further down to the infrastructure users’ affiliation has to be verified based on *some* policies. As result a series of attributes will be passed to authorisation service.

1. **Authorisation** – I would propose this to be based on UNITY AAI (<http://www.unity-idm.eu/>, very soon <https://unity.egi.eu/>. )

UNITY System knows how to talk with authentication services listed in the point one and understand attributes passed to it via affiliation check.

1. Now the fun part begins ;-)
	1. The AAI mechanism between ICS and TCSs is based on X.509 robot certificate.
	2. The ICS creates a proxy from a robot certificate and includes the user common Id in the proxy DN. The so created proxy is a per-user sub-proxy (PUSP, see [here](https://wiki.egi.eu/wiki/Fedcloud-tf%3AWorkGroups%3AFederated_AAI%3Aper-user_sub-proxy) for more details).
	3. The ICS forward the user request to the TCSs on behalf of the user
2. We will adopt an online CA when available:
	1. An online CA should be available in the EGI ecosystem in medium-long term.
	2. The online CA will replace the robot certificates.
3. TCSs recognise the proxy/token with attributes and perform requested action if the attributes allow it. :
	1. The TCS check the user’s permission on the file
	2. we should embed authorization information/attributes on the proxy.
4. Accounting data are sent back together with requested results (also not a part of this scenario but crucial for EPOS):
	1. The TCSs can store the user common Id embedded in the proxy for accounting aims

## 4.2 Extended use case, computational oriented

This use case is set up in the computational seismology domain.

It relies upon the MISFIT application. By means of such application a user can compare synthetic seismograms generated from HPC simulations results with real observations.

Although more complex than the previous basic use case, because it includes a) HPC centers, b) 3rd party applications, c) access to data residing at TCS level, this use case relies on already existing software solutions and technologies provided in the framework of the VERCE and EUDAT projects. Therefore, the amount of work needed to implement it is reasonable and fully concentrated on the orchestration of authorized access to federated resources.

### Prerequisites

1. Support of X.509 VOMS for data access and computation.
2. The VM should support EGI job management middleware (Globus)
3. The EGI system instantiates the environment and contextualizes it by making specific configurations (account, network, firewall, etc.). The authentication token of the user is copied within the virtual environment
4. While starting, the virtual environment uses the user's authentication token to remotely mount an EUDAT storage space either via fuse (irods-fuse) or gridftp (gfal?)
5. **Data should be fetched from EUDAT to the virtual environment for performance and locality reasons**
6. intermediate results may be shared with the public via the b2drop service

### User workflow

*Phase one, EIDA data staging,*

(this phase present all the technical hurdles to be tackled, which will immediately enable also phase two and three).

1. The user produces a X.509 GSI proxies via the EPOS science gateway
2. From the EPOS Science Gateway, the user discover and select the data to be downloaded
3. The user submits it to the EGI VM, which access the EIDA archives and store the data in a data management system provided by EUDAT (the EPOS Science Gateway manages the submission and provides to the EGI VM his/her proxy for authentication and authorisation).

The user interacts with the metadata to search and visualise the results.

1. The download process generates metadata, which are stored and made available as data-fabric information via the EPOS Science Gateway.

*Phase two, preprocessing of raw data and synthetics.*

1. User prepares the preprocessing workflow that will ingest raw and synthetics via the EPOS Science Gateway and selects which simulation and pre-staged raw data to use from the data-fabric catalog (provenance).

*Phase three, MISFIT workflow is applied on the preprocessed data.*

1. User prepares the MISFIT workflow via the EPOS Science Gateway and select which preprocessed data to use from the data-fabric catalog (provenance) describing the preprocessing runs
2. WHAT DOES THE USER RECEIVES, THEN??

### System workflow

*Phase one, EIDA data staging*

1. 1. The system authenticates the user on the EGI resources (VOMS)
2. 1. The system configures a data stage-in workflow
3. 1. EGI VM access the EIDA archives and store the data in a data management system provided by EUDAT
	2. The EPOS Science Gateway manages the submission and provides to the EGI VM his/her proxy for authentication and authorisation.
4. 1. The download process generates metadata which are stored and made available as data-fabric information via the EPOS Science Gateway

*Phase two, preprocessing of raw data and synthetics*

1. Synthetics have to be accessible through the EGI VM resource
2. Access is regulated via X.509 certificate

*Phase three, MISFIT workflow is applied on the preprocessed data*

1. Preprocessed raw data and and synthetics have to be accessible to the EGI VM
2. Access is regulated via X.509 certificate