EPOS: European Plate Observing System: Challenges being addressed

Keith G. Jeffery
Keith G. Jeffery Consultants
Faringdon, UK
Email: keith.jeffery@keithgjefferyconsultants.co.uk

Daniele Bailo
EPOS-ERIC Office
Istituto Nazionale di Geofisica e Vulcanologia
Rome, Italy
Email: daniele.bailo@ingv.it

Kuvvet Atakan
Department of Earth Science
University of Bergen
Bergen, Norway
Email: kuvvet.atakan@uib.no

Matt Harrison
Director Informatics
British Geological Survey
Keyworth, UK
Email: mharr@bgs.ac.uk

***Abstract*—The European plate observing system (EPOS) addresses the problem of homogeneous access to heterogeneous digital assets in geoscience of the European tectonic plate. Such access opens new research opportunities. Previous attempts have been limited in scope and required much human intervention. EPOS adopts an advanced Information and Communication Technologies (ICT) architecture driven by a catalog of rich metadata. The novel architecture together with challenges encountered and solutions adopted are presented.**

Keywords-geoscience; information; metadata; CERIF; distributed databases; research infrastructures

I. INTRODUCTION

This paper is an extended and improved version of that presented at the GeoProcessing 2019 conference [1] and details the current challenges being addressed.
First, we introduce the challenges that have faced the EPOS project and cover briefly previous relevant work.

A. Overview

Information pertaining to geoscience in Europe is heterogeneous in language, structure, semantics, granularity, content precision and accuracy, method of collection and more. However, there is an increasing demand for access to and utilisation of this information for decision-making in industry and government policy. EPOS is providing a mechanism for homogeneous access to - and comfortable utilisation of - this base of rich heterogeneous assets.

EPOS may be considered a journey. During the EPOS Preparatory Project (EPOS-PP) domain communities discovered their commonality and differences and – particularly - their digital assets offered as Thematic Core Services (TCSs). This process was lengthy, requiring much interaction to understand similarities and differences including in the use of language to describe requirements and offered assets. The whole process was facilitated by the EPOS ICT team. The assets were documented in a database, which demonstrated clearly (a) that considerable assets existed (more than 400); (b) that the organizations (covering more than 250 research infrastructures (RIs)) owning the digital assetswere willing to make them available (sometimes subject to conditions); (c) that there was overlap of assets between some communities; (d) that multidisciplinary geoscience could be achieved by providing appropriate interoperation mechanisms to make the assets available to all. An extensive review of possible architectural solutions across many sectors of research, government and industry was conducted but none satisfied the requirements. A novel, leading-edge architecture was proposed, discussed and agreed among the TCSs and the ICT team. This was then implemented as a prototype to demonstrate that, indeed, interoperation across heterogeneous communities and their assets could be achieved.

The task of the EPOS Implementation Project (EPOS-IP) is to build a geoscience environment (including governance, legal, financial, training and social aspects as well as technical ICT contributions) for the community. This Version 1.0 of the EPOS platform will then be maintained and extended by the EPOS European Research Infrastructure Consortium (EPOS-ERIC), the legal body set up by the supporting Member States providing greater sustainability for maintenance, coordination and access into the future.

There are currently 10 different TCS communities (with an additional two pending approval) with distinct and variable but complementary coverage over the entire

spectrum of solid Earth sciences. While some of the TCSs are discipline specific such as seismology, geodesy, geomagnetism, geology, others are more cross-disciplinary in their origin such as near-fault observatories, volcano observations, satellite observations of geohazards, anthropogenic hazards, multi-scale laboratories and geo-energy test-beds for low-carbon energy. Many of the assets are based on measurements by sensors or laboratory equipment covering many aspects of physics and chemistry. TCSs have variable histories of developments where some have longer history (>100 years) and hence are more mature than the others. They have established their own distinctways of working, data and software specifications. They have local domain-specific standards (although some are International or European) and constraints especially relating to their interoperation with other International organisations in their specific domain. A critical issue is the harmonisation of the descriptions of the TCSs' assets from their own local metadata standards (currently 17 different standards) as a single rich canonical metadata format with formal syntax (structure) and declared semantics (meaning of terms used). The intention is to assist interoperation of the TCSs assets within and between communities by means of the Integrated Core Services (ICS) – including the rich metadata catalog – which forms the entry-point to EPOS and the view over the EPOS assets made available withinthe TCSs.

The key requirements are as follows:

1. Minimal interference with existing communities' operations and developments including IT;
2. Easy-to-use user interface;
3. Access to assets through a metadata catalog: initially services but progressively also datasets, workflows, software modules; computational facilities, instruments/sensors all with associated organisational information including persons in roles such as experts and service managers;
4. Progressive assistance in composing workflows of services, software and data to deploy on e-Infrastructures to achieve research infrastructure user objectives.

B. Interoperability Challenge

EPOS comprises 10 communities of users characterised by domain of interest (TCSs), which supply the metadata describing the assets to the ICS. These communities have varying levels of expertise in the use of ICT for their scientific domain. The processing techniques used vary from domain to domain. With differing domains, the data models used for data collection and processing, and the metadata associated with associated services, equipment and that data, vary greatly. Across many domains geo-coordinates (including both space and time) are common, but not necessarily using the same coordinate system not standard for representation. Similarly, there are multiple metadata standards used for descriptive keywords and other attributes.

The software used for processing in each community is different, although there is some commonality, e.g., where several communities use satellite imagery. The data processing methods – from validating raw data, summarising, analytics, simulation and visualisation – varies from community to community. The more advanced communities have sophisticated workflows integrating data and processing with advanced computing facilities addressing key scientific challenges with big-data analyses and modelling. However, this is a fast-changing field and while workflows used systems like Taverna [2] in the past, the current favourite is Jupyter Notebooks [3], [4], [5]. Similarly, previous use of high-performance computers under the PRACE [6] umbrella is changing to use of commercial Cloud Computing services (such as Amazon) or EOSC (European Open Science Cloud) [7].

Most of the domains have organised computing and observational (sensor-networks) infrastructure for their purposes at institutional, national and trans-European levels. However, additionally it may be necessary to utilise supercomputing facilities, which require procurement or agreements for use as well as mechanisms to deploy the processing workflow. Progressively, EPOS is working more closely with European Open Science Cloud (EOSC) to provide such facilities, although the EPOS architecture is designed to be independent of e-Infrastructure.

e-Is (e-Infrastructures) continue to provide a level of services common to – and used by – many Research Infrastructures (RIs) and other research environments. The major e-Infrastructures of relevance to EPOS-IP are:

1. GEANT: the academic network in Europe, which brings together the national computational networks [8];
2. EGI: a foundation and organisation providing infrastructure computing and data facilities for research [9];
3. EUDAT an EC-funded project to provide infrastructure services for datasets including curation, discovery [10];
4. PRACE: a network providing resources on supercomputers throughout Europe [6];
5. EOSC: the European Open Science Cloud, which aims to provide infrastructure services for research with the first pilot project starting in January 2017 [7] and subsequently the EOSC-Hub, which is soliciting services;
6. OpenAIRE: an EC-funded project to provide metadata to access research publications and – started recently – related datasets [11].

Participant organisations in EPOS have been involved to a greater or lesser extent in all of these activities. In particular EPOS TCSs (with support from the ICS team) have been conducting pilot projects with EGI, PRACE and EUDAT and EPOS is involved in the EOSC pilot.

The level of expertise in both the science and the use of IT varies from community to community. There has been quite some education effort from the central IT team towards the domain communities to explain current computing techniques – especially for cross-domain interoperability, which previously had not been a consideration.

C. Previous Work

EPOS provides an original approach to the provision of homogeneous access over heterogeneous digital assets. Previous work has been within a limited domain (where standards for assets and their metadata may be consensual thus reducing heterogeneity) and involving much manual intervention with associated costs and potential errors. An early attempt for geoscience information was Filematch [12], which exhibited those problems. NASA has a Common Metadata Repository (CMM). In 2013 NASA decided it could not persuade every data provider to use ISO19115 so developed the Unified Metadata Model (UMM) [13] to and from which other metadata standards are converted. This follows the approach used in EPOS already and provides some assurance of the direction being taken. The Open Geoscience Consortium (OGC) has produced a series of standards. GeoNetwork [14] has established a suite of software based around the OGC ISO19115 metadata standard; however, despite its open nature this software 'locks in' the developer to a particular way of processing and does not assist in the composition and deployment of workflows and the metadata is insufficiently rich for automated processing. Some major projects run parallel to EPOS: EarthCube [15] is a collection of projects providing designs and tools for geoscience including interoperability in USA, which investigated the brokering approach – encountering the 'explosion problem' of many bilateral brokers and is now following a metadata-driven brokering mechanism like that used in EPOS, which reduces the number of convertors for metadata from (n(n-1)) to n; Auscope [16] is a set of related programmes in Australia with one (AuScope GRID) providing access to assets and using ISO19115 as the metadata standard with the deficiencies mentioned above; GEOSS [17] is developing interoperation through a system or systems approach, which naturally requires many bilateral interfaces to be maintained with consequent difficulties and maintenance costs as systems evolve.

Thus, the EPOS solution overcomes the major problems associated with previous or parallel work namely: many-to-many interfaces between software brokers or systems and insufficiently rich metadata for automation while enabling interoperability across multiple asset sources.

On October the 30th 2018, the European Commission granted the legal status of European Research Infrastructure Consortium (ERIC) to EPOS, which was already promoted as a landmark in the ESFRI 2018 Roadmap.

The rest of the paper is organized as follows: Section II describes the architecture; Section III discusses the importance of metadata and Section IV discusses the major challenges faced currently and progress towards solutions and Section V gives the current state and outlook.

II. ARCHITECTURE

The ICT architecture of EPOS is designed to facilitate the research community and others in discovering and utilizing through the ICS the assets provided by the TCS communities.

A. Introduction

In order to provide end-users with homogeneous access to services and multidisciplinary data collected by monitoring infrastructures and experimental facilities (and to software, processing and visualization tools as well) a complex scalable and reliable architecture is required. A snapshot of the architecture is outlined in Figure 1. It includes three main layers:

Integrated Core Services – ICS, the core component designed and run by EPOS; this is the place where the integration of data and services provided by the TCS, Community Layer occurs. 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 Distributed Services (ICS-D), whose goal is to provide resources (e.g., computational, visualisation).

Thematic Core Services – TCS, made up of pan European e-Infrastructures, which disseminate data and services of a single discipline (e.g., seismology with ORFEUS/EIDA).

National Research Infrastructures – NRI, made up of RIs providing data and services,

Starting from the latter, NRI represent the wealth of assets provided by national or regional institutions or consortia, and are referred to as DDSS, i.e., Data, Data-products, Software and Services. The asset descriptions were collected first as DDSS in the DDSS master table (stored in Excel), which also records the state of maturity and management parameters. This is now being replaced progressively by the so-called Granularity Database

(GRDB), which records the same information as the DDSS master table but using the same metadata standard as that of the ICS-C catalog (described below in Section III) for ease of managing the process of approving a DDSS for inclusion in the ICS-C metadata catalog. The GRDB DDSS records are harvested as metadata for population of the EPOS ICS-C catalog.

TCSs enable the integration of data and services from specific scientific communities. The architecture of the services provided by the individual communities is not prescribed, what is required is that the metadata describing the data and services available is in a form that can be consumed by the ICS, allowing the ICS to integrate with those services and data (Figure 1).

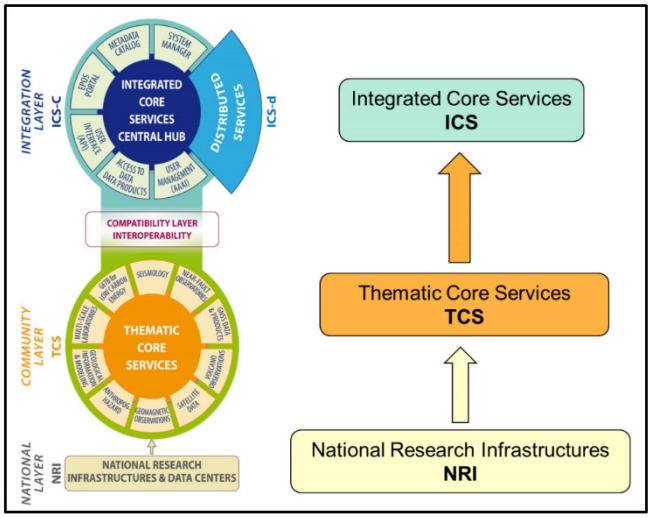

Figure 1. EPOS Architecture

B. ICS

The EPOS-ICS provides the entrypoint to the EPOS environment. The ICS consists of the ICS-C and distributed computational resources including also processing and visualisation services (ICS-D) of which a specialization is Computational Earth Science (CES). ICS-C provides a catalog of, and access to, the assets of the TCSs. It also provides access to e-Infrastructures (e-Is)as ICS-Ds upon which (parts of) workflows are deployed (other parts may be deployed within the computing capabilities of RIs within EPOS). EPOS has been involved in projects with e-Is to gain joint understanding of the interfaces and capabilities ready for deployment from ICS-C. EPOS has also been involved in the VRE4EICproject [18] (and cooperating with EVER-EST [19]) to ensure convergent evolution of the EPOS ICS-C user interface and APIs for programmatic access with the developing Virtual Research Environments (VREs). EPOS partners are also participating inthe recently approved ENVRIFAIR [20] project, which will assist in building linkages between EPOS ICS-C and European Open Science Cloud (EOSC) (Figure 2).

The linkage between ICS-C on the one hand and the e-Is and TCS local computing resources and assets on the other is through ICS-Ds, which will be constructed as a workflow in the ICS-C and managed in the deployment phase. The workflow for the deployment (which may be a simple file download or a complex set of services including analytics and visualisation) will be generated within the ICS-C by interaction with the users. The workflow will be checked by the end-user before deployment. However, the detailed content/capability of the assets might not be known, e.g., the dataset may not contain the relevant information despite its metadata description, or the software may not execute as the user expects despite the metadata description.

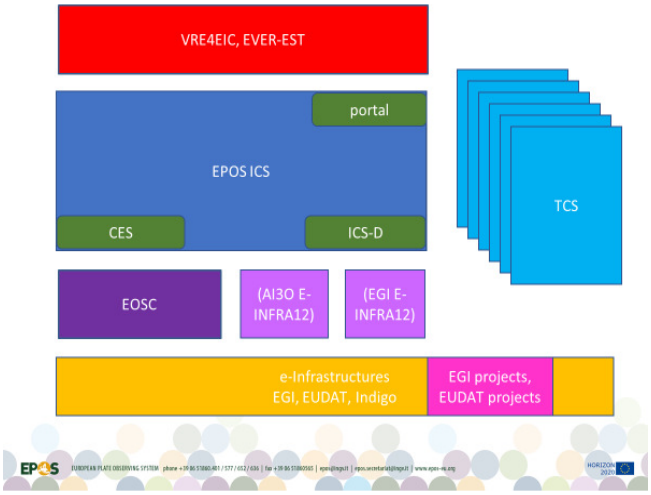

Figure 2. EPOS Positioning

The execution of the deployment is monitored and execution information is returned to the end-user. The workflow may be deployed in one of two ways: (a) directly with no user interaction during execution of the deployment; (b) step-by-step with user interaction (so-called computational steering) between each step. Deployments of type (a) will have better optimisation (for performance) and security but could possibly execute a workflow, the components of which do not behave as the user expects. Deployments of type (b) lack optimisation but allow the user to stop the workflow deployment at any step, examine the results and – if not as expected – reorganise the workflow (by changing components) to meet more closely the requirement.

The ICS represents the infrastructure consisting of services that will allow access to multidisciplinary resources provided by the TCS. These will include data

and data products as well as synthetic data from simulations, processing, and visualization tools.

C. ICS-C

The ICS-C consists of multiple logical areas of functionality, these include the Graphic User Interface (GUI), web-API, metadata catalogue, user management etc. A micro-service architecture has been adopted of the ICS-C, where each (micro) services is atomic and dedicated to a specific class of tasks. The ICS-C is where the integration of other services from ICS-D and TCS takes place. The architectural constraints for the ICS-D are elaborated as a metadata model within the ICS-C CERIF (Common European Research Information Format) [21] catalog and are being implemented.

The ICS-C System is the main system that manages the integration of DDSS from the communities. On top of such a system, a Graphic User Interface (GUI) enables the user to search, discover and integrate data in a user-friendly way.

The EPOS ICS-C system architecture (Figure 3) was designed and developed with the aim of integrating data and services provided by TCS. In order to a) enable the system to run in a distributed environment, b) guarantee up-to-date technological upgrades by adopting a software-independent approach, c) proper scaling of specific system functionalities, the chosen architecture followed a microservices paradigm.

The Microservices architecture approach envisages small atomic services dedicated to the execution of a specific class of tasks, which have high reliability [22], [23]. Such architecture replaces the monolith with a distributed system of lightweight, narrowly focused, independent services. In order to implement the microservices paradigm, Docker Containers technology was used [24]. It enables complete isolation of independent software applications running in a shared environment. In particular, each microservice is developed in the Java language and performs a simple task, as atomic as possible. The communication between microservices is done via messages received and sent on a queueing system, in this case RabbitMQ [25]. As a result, a chain of microservices processes the requests.

The current architecture includes an Authentication, Authorisation, Accounting Infrastructure (AAAI) module. This has been implemented using UNITY [26] and has involved close cooperation with CYFRONET. Since May 2018 this has formed the basis of anintegrated authentication system for academic communities Authorisation is more complex and depends on rules agreed with the TCS (within the context of the financial, legal and governance traversal workpackages of EPOS-IP) for each of their assets and included further metadata elements into the CERIF catalog to control such authorisation. AAAI will be continuously evolved and updated to ensure appropriate security, privacy and governance. Related to this, the GUI now provides a user notification pointing to a legal disclaimer for the EPOS system, terms and conditions and acceptance of cookies.

A major requirement of the system, after asset discovery, is the construction of workflows that can be used to access / process data. This has implications for the entire software stack; visually designing the workflows, managing and persisting inputs and outputs, scheduling and execution of processes, access to metadata, access to data and service from the TCS. The topic as whole required significant analysis of requirements and available technologies. Working in cooperation with the VRE4EIC project we have the basic components for (a) a general workflow manager interface; (b) interfaces to specific workflow managers such as Taverna [2].

Beyond simple map visualisations that consume web map services the ICS-C user interface may be required to support additional types of visualisation. This set of supported visualisation types and associated data formats is being confirmed with the TCS representatives through a series of ongoing workshops as it will not be practical to support all formats of data for all types of visualisation.

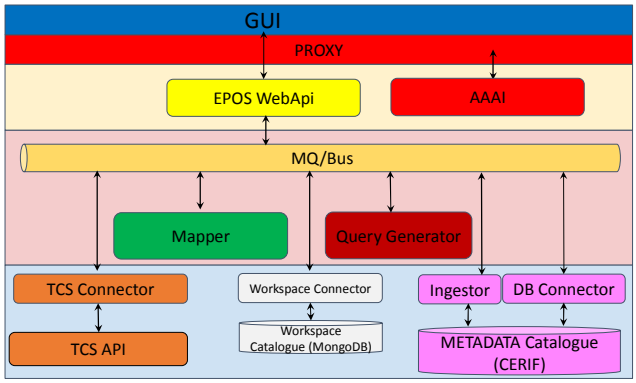

Figure 3. ICS-C Architecture

D. ICS-D

The distributed services offered by the ICS-D facet of the architecture ties-in with the workflow management, as the distributed services in question - beyond just being discoverable - are likely candidates for inclusion in processing workflows. A specification of the metadata elements required for ICS-D has been produced, is under review and forms part of the architecture. ICS-D will appear to the workflow, or to the end-user, as a service accessed through an API. However, the choice of which ICS-D to use and the deployment of a workflow across one or more ICS-Ds requires optimisation middleware.

Results from the PaaSage project [27] are relevant and the concurrent MELODIC project [28] offers optimisation including that based on dataset placement and latency. Further refinement of requirements and the architectural interfaces continues.

III. METADATA

Metadata is the key to discover and utilise the heterogeneous assets of EPOS in a homogeneous way thus facilitating cross-domain, interoperable science.

A. Introduction

The metadata catalogue is the key technology that enables the system to manage and orchestrate all resources required to satisfy a user request. By using metadata, the ICS-C can discover data or other digital objects requested by a user, contextualise them (for relevance and quality) access them, send them to a processing facility (or move the code to facility holding the data) depending on the constructed workflow, and perform other tasks. The catalogue contains: (i) technical specification to enable autonomic ICS access to TCS discovery and access services, (ii) metadata associated with the digital object with direct link to it, (iii) information about users, resources, software, and services other than data services (e.g., rock mechanics, geochemical analysis, visualization, processing). The data model used for the catalogue is CERIF.

Metadata describing the TCS DDSS are stored using the CERIF data model, which differs from most metadata standards in that it (1) separates base entities from linking entities thus providing a fully connected graph structure; (2) using the same syntax, stores the semantics associated with values of attributes both for base entities (to ensure valid attribute values are recorded for instances, e.g., ISO country codes) and for linking entities (for role of the relationship), which also store the temporal duration of the validity of the linkage. This provides great power and flexibility. CERIF also (as a superset) can interoperate with widely adopted metadata formats such as DC (Dublin Core) [29], DCAT (Data Catalogue Vocabulary) [30], CKAN (Comprehensive Knowledge Archive Framework) [31], INSPIRE (the EC version of ISO 19115 for geospatial data) [32] and others using convertors developed as required to meet the metadata mappings achieved between each of the above standards and CERIF. The metadata catalogue also manages the semantics, in order to provide the meaning of the instance attribute values. The structure of base entities and linking entities used for metadata instances is also used for the semantic layer of CERIF; the base entities containing lexical entries and the linking entities maintaining the relationships between them allowing a full ontology graph structure including not only subset and superset terms but also equivalent terms (especially useful for multilinguality) and any other role-based logical relationship between terms.

The use of CERIF provides automatically:

(a) The ability for discovery, contextualization and (re-)use of assets according to the FAIR principles [33];
(b) A clear separation of base entities (things) from link entities (relationships);
(c) Formal syntax and declared semantics;
(d) A semantic layer also with the base/link structure allowing crosswalks between semantic terminology spaces;
(e) Conversion to/from other common metadata formats;
(f) Built-in provenance information because of the timestamped role-based links;
(g) Curation facilities because of being able to manage versions, replicates and partitions of digital objects using the base/link structure.

The catalog is constantly evolving with the addition of new assets (such as services, datasets) but also increasingly rich metadata as the TCSs improve their metadata collection to enable more autonomic processing.

B. TCS Metadata

The process of populating the catalog is crucial in the EPOS vision. Indeed, populating the catalog means to make available all the information needed by an end user to perform queries, data integration, visualisation and other functionalities provided by the system.

Greater interaction with TCS communities to ensure that their metadata, data and services are available for harvesting in the appropriate format and to populate the CERIF data model has been achieved and will be continued.

C. ICS Metadata

In order to manage all the information needed to satisfy user requests, all metadata describing the TCS Data, Datasets, Software and Services is stored into the EPOS ICS, internal catalog, based on the aforementioned CERIF model, which differs from most metadata standards used by various scientific communities in that it is much richer in syntax (structure) and semantics (meaning).

For this reason, EPOS ICS has sought to communicate to the TCS communities the core elements of metadata required to facilitate the ICS through the EPOS Metadata Baseline. This baseline can be considered as an intermediate layer that facilitates the conversion from the community metadata standards such as ISO19115/19, DCAT, Dublin Core, INSPIRE etc. describing the DDSS

elements and not the index or detailed scientific data (Figure 4).

The EPOS baseline presents a minimum set of common metadata elements required to operate the ICS taking into consideration the heterogeneity of the many TCSs involved in EPOS. It has been implemented as an application profile using an extension of the DCAT standard, namely the EPOS-DCAT-AP [34]. It is possible to extend this baseline to accommodate extra metadata elements where it is deemed that those metadata elements are critical in describing and delivering the data services for any given community. Indeed, this has happened when the original EPOS-DCAT-AP was found to be inadequate and a new version with richer metadata was designed and implemented.

The metadata to be obtained from the EPOS TCSs as described in the baseline document (and any other agreed elements) are mapped to the EPOS ICS CERIF catalog. The process of converting metadata acquired from the EPOS TCS to CERIF isdone in consultation with each TCS as to what metadata they have available and harvesting mechanisms.

The various TCS nodes have APIs or other mechanisms to expose the metadata describing the available DDSS in a TCS specific metadata standard that contains the elements outlined in the EPOS baseline documents better described in the following sections. It also requires ICS APIs (wrappers) to map and store this in the ICS metadata catalogue, CERIF. These APIs and the corresponding ICS convertors collectively form the "interoperability layer" in EPOS, which is the link between the TCSs and the ICS.

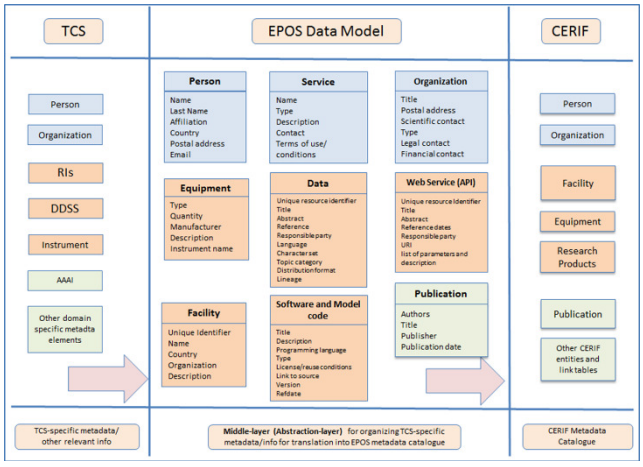

Figure 4. EPOS Metadata Baseline

D. DDSS and Granularity Database

As a part of the requirements and use cases collection (RUC) from the TCSs, a specific list was prepared to include all data, data products, software and services (DDSS). This DDSS Master Table was used as a mechanism to update the RUC information as well as providing a mechanism for accessing more detailed IT technical information for the development of the ICS Central Hub (ICS-C). The DDSS Master Table was also used for extracting the level of maturity of the various DDSS elements in each TCS as well as providing a summary of the status of the TCS preparations for the ICS integration and interoperability. The current version of the DDSS Master Table consists of 368 DDSS elements, where 201 of these already exist and are declared by TCSs to be ready for implementation. The remaining DDSS elements required more time to harmonize the internal standards, prepare an adequate metadata structure and so are available for implementation soon. In total, 21 different harmonization groups (HGs) are established within the EPOS-IP project to help organizing the harmonization issues in a structured way. TCSs are preparing individual TCS Roadmaps, which will describe the development and implementation plans of the remaining DDSS elements including a time-line and resource allocations. In addition, user feedback groups (UFGs) are being established in order to give constant and structured feedback during the implementation process of the TCS-ICS integration and the development of the ICS.

The DDSS Master Table was constantly being updated as new information from the TCS WPs arrive. The older versions are also kept in the archive for future reference. The DDSS master table is being transformed to the GRDB (granularity database) because of the problems of referential and functional integrity using a spreadsheet; relational technology provides appropriate constraints to ensure integrity. As such, the GRDB represents a structured way of requirements and use cases collection (RUC) from the TCS communities. Updates or new entries to GRDB can be done either using a dedicated GUI or in an automated manner.

The TCS requirements and use cases (RUC) collection process was designed carefully, taking into account the amount and complexity of the information involved in all 10 different TCSs. An increasingly detailed RUC collection process is formulated and explained through dedicated guidelines and interview templates. A roadmap for the ICS-TCS interactions for the RUC collection process was prepared for this purpose and distributed to all TCSs.

In this approach, a five-step procedure is applied involving the following:

- Step 1: First round of RUC collection for mapping the TCS assets;
- Step 2: Second round of RUC collection for identifying TCS priorities;
- Step 3: ICS-TCS Integration Workshop for building a common understanding for metadata;
- Step 4: Third round of RUC collection for refined descriptions before implementation;
- Step 5: Implementation of RUC to the CERIF metadata.

Planning for the requirements and use cases (RUC) elicitation process started with the pre-project meeting held during the period July 8-9 2015 at the BGS (British Geological Survey) facilities in Nottingham, UK. The first version of the guidelines level-1 for the ICS-TCS integration was prepared soon after this meeting and was distributed to the TCS leaders and the relevant IT-contacts. A second, more detailed guidelines level-2 was prepared in September 2015 and distributed in the EPOS-IP project kick-off meeting held in Rome, Italy, during the period October 5-7 2015. Prior to the kick-off meeting, a preliminary collection of the RUC was requested from each TCS, which was then presented during the meeting.

In parallel with the guidelines for the ICS-TCS Integration, a dedicated RUC interview template level-1 was preparedto be used during the first site visits to the TCSs. The site visits were conducted during the time period between November 2015 and March 2016. All four steps are now completed, whereas step 5 with metadata implementation has started in January 2017 and is ongoing.

Work is almost complete in converting the DDSS tables (in Excel) to the GRDB using Postgres. This will (a) facilitate finding particular DDSS elements, eliminating duplicates and checking the progress of getting DDSS elements into metadata format; (b) actually harvesting to the metadata catalog.

IV. CURRENT CHALLENGES

This section lists the current challenges being addressed, beyond the system as described in [1].

A. Introduction

A project as large in terms of organisations, persons and assets involved and as complex in terms of governance, funding and technology required, necessarily faced many challenges. Some of the key challenges are discussed.

B. Metadata Conversion

As discussed in Section III, the use of a canonical rich metadata format is key to providing homogeneous access to the heterogeneous assets within EPOS. Reaching the state of all assets recorded in this standard posed some challenges. These are outlined below.

1) Heterogeneity

However, the multiple metadata ‘standards’ used widely within the various EPOS communities – and in some cases used by those communities within an international context for exchange of data – needed to be respected while converting to the canonical rich metadata standard CERIF. This conversion was achieved by much discussion between each TCS community and the ICS ICT team. The discussion involved understanding not only the metadata model being used (which usually was well-documented) but also how it was used – with which interpretation of the ‘rules’ of the model. As well as the heterogeneity in the ‘standards’ used, there was also heterogeneity in its interpretation, even of the same ‘standard’.

2) Complexity

CERIF provides a rich metadata model. Mathematically it is a fully connected graph. The metadata ‘standards’ used by the TCS communities were – in general – simple, consisting of records not unlike a library catalog card with attributes related to an asset such as a service or dataset. These attributes commonly included persons and organisations, which could be multiple and were not functionally dependent on the asset being described; this meant that the TCS metadata records did not have referential and functional integrity. However, the TCS communities were familiar with their own ‘standard’ and found difficulty in understanding (a) the concept of integrity to ensure validity of the metadata; (b) the need for a fully connected graph structure to represent more accurately the real world. As described in Section III, this problem was overcome by using a simplified intermediate format (EPOS-DCAT-AP), which – stored in RDF (Resource Description Framework) - acted as a ‘bridge’ between the simple metadata structures of the TCSs and the richness of CERIF.

C. Legal, Governance and AAAI Aspects

The overall intention of EPOS is to make assets findable, accessible, interoperable and reusable in an open environment and toll-free to not-for-profit users. However, it was necessary to introduce some technical ICT features to accommodate legal, governance and AAAI aspects.

1) Terms and Conditions of Use

A conditions of use document was produced and made accessible from the ‘landing page’ (the screen first encountered when accessing EPOS) with a requirement that a user should accept the Terms and Conditions.

2) Disclaimer

Similarly, a disclaimer document was produced and made accessible from the 'landing page' with a requirement that a user should accept the Terms and Conditions.

3) Cookies

Also, on the 'landing page' there is a requirement for the user to accept (or not) the use of cookies in EPOS.

4) AAAI – Authentication

There is a need to authenticate users (i.e., ensure the user has credentials to assure that they are who they claim to be) for several reasons. (a) it provides security against individuals accessing the system with malicious intent; (b) it allows audit and provenance trails to be related to a person for several purposes: to provide records to demonstrate compliance with GDPR (General Data Protection Regulation); to allow reproduction of the scientific pathway to corroborate research results; to improve user interaction by suggesting (based on past usage) assets to be used. EPOS aligns with current leading-edge work in this area using authentication agents such as EduGAIN [35] and also tracks the ongoing work within the European AARC2 project [36].

5) AAAI-Authorisation

Once a user is authenticated, he/she may be authorized (by some other authority) to access assets. The access may be restricted by role of the user, by time interval, by the process intended (e.g., read, execute, modify, delete) as well as by collection of assets or individual asset. The authorization system is currently being discussed with the TCS representatives since (a) it requires collection of more metadata for the assets, persons and organisations; (b) it requires appropriate access control program code to be provided.

D. Use of DoI (Digital Object Identifier)

A problem for a particular collection of assets is the use of DoI. The DoI system works by dereferencing the DoI to a landing page, which contains text describing the asset and a URL, which dereferences to the asset itself. The concept is based on human interaction, the human reads the landing page text and decides whether to access the asset.

In contrast, the EPOS ICS-C is based around the concept that the user queries the metadata catalog for assets that – satisfying the query - are relevant and of sufficient qualityto allow automated access - and then accesses them directly.

Two solutions are being worked upon: (a) for those DoI-based collections, which have a well-structured landing page template to use MIME types to access the URL pointing directly to the asset, thus 'bypassing' the step of a human reading the landing page (although the lack of rich metadata in the metadata catalog may well mean that relevant assets are not recalled by the query); (b) where the landing page text is well-structured, converting the metadata text of the landing page to a CERIF record in the metadata catalog together with the asset URL thus rendering the landing page redundant.

E. Complexity of the GUI (Graphical User Interface)

Different TCSs have different ways of finding, accessing, interoperating and re-using assets. The design challenge was to find a common process structure with step sequences (including cycling back to previous steps) to accommodate these different requirements. In turn this made the design of the GUI more complex since different users wished to traverse the process steps in different ways. At workshops involving TCS community representatives and the ICS ICT team scientific stories were mapped to use cases, and these were used to define the GUI requirements.

The complexity arises because users may wish to confirm their choice of a single asset by seeing it visually – on a map or chart – before deciding whether to add the metadata for that asset to the workspace that they are constructing during the session. Furthermore, they may wish to change the parameters of the asset – especially of a service – and re-visualise. On the other hand, some users wish to see the assets represented by metadata in the workspace visualized as a 'build-up' with each one overlaid on the other. Thereafter they may wish to change the parameters of one or more assets before composing a workflow, which involves cycling back to visualization of single assets before checking again the 'build-up' of visualisations for all the assets represented by metadata in the workspace.

Different possibilities are being tested with representative TCS users to determine which options should be implemented in the operational system due to be released end-September 2019.

F. Intersection of AAAI with GUI

Another challenging factor is the question of when to demand that a user is authenticated. Some (few) users wish completely anonymous, open, toll-free access. This is clearly not possible for legal and governance reasons; for example, the potential for liability litigation or the potential use of a large amount of supercomputing resources without prior authorization.

There is, however, an argument for a user being able to query the metadata catalog and visualize individual selected assets to see if they suit his/her requirements

before logging in / authenticating prior to composing or deploying a workflow. Furthermore, this approach leaves the TCS communities free to control authorization of asset usage since login and authentication takes place before access to the assets with authorization. However, this approach leaves metadata catalog access open to a liability challenge (since the user will not yet have accepted the disclaimer) and also may contravene GDPR (since the metadata includes information about persons - such as the owner of an asset or the manager of an asset).

The safest approach is to demand login / authentication at session start. This ensures not only security and legal/governance compliance but also initiates appropriate audit and provenance recording. The counter-argument is that immediate login/authentication may be a barrier to use of the system for some users.The ICS ICT team is currently discussing these options with both EPOS governance structures and TCS users to find the appropriate design that can be implemented.

V. CONCLUSION AND FUTURE WORK

The European plate observing system (EPOS) is addressing the challenges of accessing heterogeneity in a homogenous way by building an integration node called Integrated Core Services. This system is metadata driven and uses the CERIF model. Currently 136 distinct DDSS accessible through 264 different web-services from the domain communities are represented by CERIF metadata in the EPOS ICS-C catalog. These services, described by the metadata, can be discovered, contextualised and utilised individually or composed into workflows and hence become interoperable. A GUI (Graphical User Interface) provides the user view onto the catalog, and it also provides a workspace to collect the metadata of the assets selected for use (Figure 5). From the workspace a workflow may be constructed and deployed.

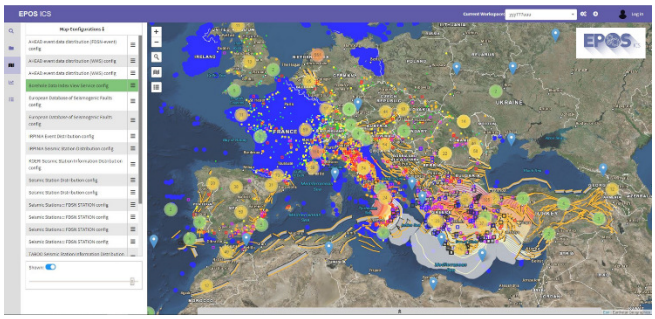

Figure 5. EPOS-ICS graphical user interface.

Future plans include:

a) Harvesting of metadata describing more assets: not only services but also datasets, software, workflows, equipment;
b) Improving the GUI to allow workflow deployment with 'fire and forget' technology or single-step with user checking and adjustment at each step;
c) Completion of the (current prototype) software to permit trans-national access to laboratory and sensor equipment;
d) Improved AAAI (Authentication, authorisation, accounting infrastructure) to give the domain users finer-grained control over access to their assets;
e) The inclusion of virtual laboratory-type interfaces (virtual research environments) allowing users access and connectivity including open-source frameworks such as Jupyter notebooks [3], which are increasingly being used in some scientific communities.

The architecture outlined and demonstrated (in successive prototypes) in EPOS-IP has found favour (not without some criticism of course – leading to agile improvements) from the user community. Furthermore, the prototype system has passed Technological Readiness Assessment procedures within the governance of the EPOS-IP project. Currently the ICS is undergoing validation tests. The first operational release is scheduled for end-September 2019. The architecture meets the requirements, it is state of the art and has a further development plan.

ACKNOWLEDGMENT

The authors acknowledge the work of the whole ICT team in EPOS reported here and the funding of the European Commission H2020 program (Grant agreement 676564) and National Funding Councils that have made this work possible.

REFERENCES

[1] K. Jeffery, D. Bailo, K. Atakan, and M. Harrison, "EPOS: European Plate Observing System," in Proc. Eleventh International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2019), pp. 79-86.
[2] Taverna: https://taverna.incubator.apache.org/2019.11.11
[3] Jupyter: https://jupyter.org/2019.11.11
[4] F. Pérez and B. Granger, "IPython: a system for interactive scientific computing," Computing in Science and Engineering, 9(3), pp. 21-29, June 2007.
[5] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, and P. Ivanov, "Jupyter Notebooks-a publishing format for reproducible computational workflows," in Proc. 20th International Conference on Electronic Publishing (ELPUB), pp. 87-90, May 2016.

[6] PRACE: http://www.prace-ri.eu/ 2019.11.11
[7] EOSC pilot: https://eoscpilot.eu/ 2019.11.11
[8] GEANT: http://www.geant.org/ 2019.11.11
[9] EGI: https://www.egi.eu/ 2019.11.11
[10] EUDAT: https://eudat.eu/ 2019.11.11
[11] OpenAIRE: https://www.openaire.eu/ 2019.11.11
[12] P. Sutterlin, K. Jeffery, and E. Gill, "Filematch: A Format for the Interchange of Computer-Based Files of Structured Data," Computers and Geosciences, Vol. 3 (1977), pp. 429-468.
[13] UMM: https://earthdata.nasa.gov/about/science-system-description/eosdis-components/common-metadata-repository/unified-metadata-model-umm 2019.11.11
[14] Geonetworkhttps://geonetwork-opensource.org/ 2019.11.11
[15] EarthCube: https://www.earthcube.org/ 2019.11.11
[16] AuScope: http://www.auscope.org.au/ 2019.11.11
[17] GEOSS: https://www.earthobservations.org/geoss.php 2019.11.11
[18] VRE4EIC: https://www.vre4eic.eu/ 2019.11.11
[19] EVEREST: https://ever-est.eu/ 2019.11.11
[20] ENVRI-FAIR website http://envri.eu/envri-fair/ 2019.11.11
[21] CERIF: https://www.eurocris.org/cerif/main-features-cerif 2019.11.11
[22] S. Newman, "Building Microservices," O'Reilly Media, Inc., February 2015, ISBN: 9781491950340.
[23] D. Namiot and M. Sneps-Sneppe, "On Microservices Architecture," International Journal of Open Information Technologies, ISSN 2307-8162, Vol. 2, No. 9, pp. 24-27, 2014.
[24] Docker: https://www.docker.com/ 2019.11.11
[25] RabbitMQ: https://www.rabbitmq.com/ 2019.11.11
[26] UNITY: http://www.unity-idm.eu 2019.11.11
[27] PaaSage: https://paasage.ercim.eu/ 2019.11.11
[28] MELODIC: melodic.cloud/ 2019.11.11
[29] DC: http://dublincore.org/documents/dces/ 2019.11.11
[30] DCAT: https://www.w3.org/TR/vocab-dcat/ 2019.11.11
[31] CKAN: https://ckan.org/ 2019.11.11
[32] INSPIRE: https://inspire.ec.europa.eu/ 2019.11.11
[33] FAIR: https://www.force11.org/grohttps://ckan.org/up/fairgroup/fairprinciples 2019.11.11
[34] EPOS-DCAT-AP on GitHub: https://github.com/epos-eu/EPOS-DCAT-AP 2019.11.11
[35] https://edugain.org/ 2019.11.11
[36] https://aarc-project.eu/ 2019.11.11