Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

13th International Conference on Current Research Information Systems, CRIS2016, 9-11 June 2016, Scotland, UK

Mapping solid earth Data and Research Infrastructures to CERIF

Daniele Bailo[a]*, Damian Ulbricht[b], Martin L. Nayembil[c], Luca Trani[d], Alessandro Spinuso[d], Keith G. Jeffery[e]

[a]*INGV - Istituto Nazinoale di Geofisica e Vulcanologia, via di Vigna Murata 605, 00143, Roma, Italy, daniele.bailo@ingv.it*
[b]*GFZ German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany, ulbricht@gfz-potsdam.de*
[c]*British Geological Survey, Nicker Hill, Keyworth,Nottingham, NG12 5GG, U, mln@bgs.ac.uk*
[d]*KNMI, Utrechtseweg 297,3731 GA, De Bilt, the Netherlands, trani@knmi.nl - spinuso@knmi.nl*
[e]*Keith Jeffery Consultant, Shrivenham, United Kingdom keith.jeffery@keithgjefferyconsultants.co.uk*

Abstract

EPOS is a Research Infrastructure plan that is undertaking the challenge of integrating data from different solid Earth disciplines and of providing a common knowledge-base for the Solid-Earth community in Europe, by implementing and managing a logically centralised catalog based on the CERIF model. The EPOS catalogue will contain the information about all the participating actors, such as Research Infrastructures, Organisations and their assets, in relationship with the people, their roles and their affilitation within the specific scientific domain. The catalogue will guarantee the discoverability of domain specific data, data products, software and services (DDSS) and enable the EPOS Integrated Core Services system to perform - on behalf of a end user – advanced operations on data as for instance processing and visualization. It will also foster the homogenisation of vocabularies, as well as supporting heterogeneous metadata. Clearly, the effort of accomodating the diversities across all the players needs to take into account of existing initiatives concerning metadata standards and institutional recommendations, trying to satisfy the EPOS requirements by incorporating and profiling more generic concepts and semantics. The paper describes the approach of the EPOS metadata working group, providing the rationale behind the integration, extension and mapping strategy to converge the EPOS metadata baseline model towards the CERIF entities, relationships and vocabularies. Special attention will be given to the outcomes of the mapping process between two elements of the EPOS baseline - Research Infrastructure and Equipment - and CERIF, by providing detailed insights and description of the two data models, of encountered issues and of proposed solutions.

Peer-review under responsibility of the Organizing Committee of CRIS2016.

Keywords: Research Infrastructure, metadata, e-infrastructure, data integration, CERIF, EPOS

* Corresponding author. Tel.: +39 0651860728; Fax: +39 06-51860507.
E-mail address: daniele.bailo@ingv.it.

Peer-review under responsibility of the Organizing Committee of CRIS2016.

1. Introduction

In the last two decades the interest in environmental data has raised incredibly in several domains. The reason for this growing interest is the awareness of the value of the environmental data and its disruptive potential in solving some of the environmental challenges identified by the EU [14], which include for instance flood and landslide risks, geo-hazards, anthropogenic hazards, climate change and others. The main question, in this scenario, is how to integrate multiple sources that provide access to data and services at National or Regional level (e.g. local data centers). Such data providers present the following characteristics: a) they are usually scattered over Europe, b) they often use community specific standards (even though initiatives like INSPIRE[†] provide common guidelines at european level), c) data, services and results of a discovery action are seldom standard, and depend upon the local technologies.

As a consequence, various actors started a number of initiatives such as the above mentioned INSPIRE, that at European governmental level provides a common framework of Implementing Rules that, when adopted at National level, ensure that "spatial data infrastructures of the Member States are compatible and usable in a Community and transboundary context". Also, domain specific projects or initatives exist at pan-european level; this is the case of: US NSF EarthCube that aims at "develop a common cyberinfrastructure for the purpose of collecting, accessing, analyzing, sharing and visualizing all forms of data and related resources "[‡] in Geoscience field; the ENVRIplus[§] initiative is dedicated to bringing together Environmental and Earth System Research to create a more coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe; the GEOSS – Global Earth Observation System of Systems [**], whose goal is to link independent Earth observation, information and processing systems to strengthen the monitoring of the state of the Earth; not to mention the marine domain with EMSO[††], ELIXIR[‡‡] in life science and many others.

In the domain of solid Earth Sciences, the European Plate Observing System (EPOS)[§§] project, now in its Implementation Phase, is building a pan-European Research Infrastructure to integrate solid Earth data, data products and services. The key challenge to achieve this integration is a proper management of metadata. In the context of EPOS the metadata challenge was tackled choosing a metadata driven brokering approach: such approach brings together the advantages of a brokering system, which provides a single homogeneous access to heterogeneous resources, and the advantages of a metadata driven system, which can easily manage all the elements involved in the processing of a user request by using the metadata associated to them. In this framework, a comprehensive metadata model, that could handle metadata describing not only data but also users, software and resources, was needed. Such a model was identified in CERIF[***].

The aim of this paper is therefore to describe how CERIF is being used in the EPOS initiative and how concepts from the solid Earth science domain were mapped to CERIF model. Building on results from previous initiatives, we will a) describe the process of mapping metadata in a context where the community is huge and scattered over Europe, b) show how the *Research Infrastructure* and *Equipment* concepts can be mapped to CERIF, c) point to the difficulties, solutions and lessons learnt in this process and d) propose some possible future develpoment and extensions that may enable CERIF to manage a high volume of users and data products from different contexts in the solid Earth science domain.

[†] http://inspire.ec.europa.eu/
[‡] http://earthcube.org/
[§] http://www.envriplus.eu/
[**] http://www.earthobservations.org/geoss.php
[††] http://www.emso-eu.org/
[‡‡] https://www.elixir-europe.org/
[§§] https://www.epos-ip.org/
[***] http://www.eurocris.org/cerif/main-features-cerif

2. What is EPOS

The European Plate Observing System (EPOS) has been designed with the vision of creating a pan-European e-Research Infrastructure for solid Earth science to support a safe and sustainable society. In accordance with this scientific vision, the EPOS mission is to integrate the diverse and advanced European Research Infrastructures for solid Earth science relying on new e-science opportunities to monitor and unravel the dynamic and complex Earth System. EPOS will enable innovative multidisciplinary research for a better understanding of the Earth's physical and chemical processes that control earthquakes, volcanic eruptions, ground instability and tsunami as well as the processes driving tectonics and Earth's surface dynamics. A detailed description of EPOS is provided by Bailo et al[1].

3. EPOS architecture

In order to contextualise the work presented in this paper, a short description of EPOS architecture is provided. Such architecture represents both a governmental and technical organization of the main actors of EPOS e-RIs. Interoperability among the systems of these actors is needed in order to provide to the end user the required functionalities, i.e. discovery, download, processing and visualization of solid Earth science related data at European level. The EPOS community is organized from top to bottom (Fig.1a) as follows:

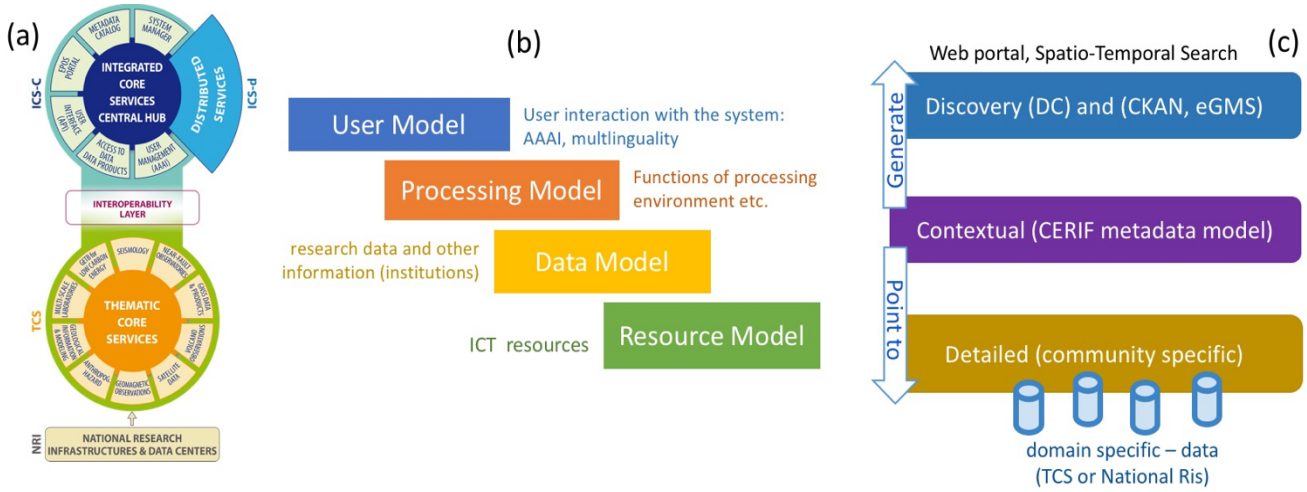

Figure 1: (a) EPOS architecture. From top to bottom: Integrated Core Services – ICS, Thematic Core Services – TCS, National Research Infrastructures – NRI; (b) the four metadata dimensions: Users, Processing, Data, Resources; (c) the three layer metadata model[5], that includes 1.*A discovery layer*, utilising the capability to generate from the underlying contextual layer – providing DC, DCAT, INSPIRE (to integrate with other existing datacentres utilizing these standards) and both CKAN and eGMS (the latter two to foster integration with government open data (data.gov) sources); 2. *A contextual layer*, using CERIF from which the discovery level metadata standards can be generated and which also points to individual datasets or services metadata in the detailed layer; 3. *A detailed metadata layer*, which includes detailed metadata standards usually domain or sub-domain specific for each kind of data (or software, computer resources or detectors/instruments) to be (co)-processed.

1. *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. Integrated Core Services are characterized by a *Central Hub (ICS-C),* whose main goal is to host the metadata catalog and orchestrate external resources (e.g. HPC), and the *Distributed Services (ICS-d),* whose goal is to provide computational and visualization resources.

2. *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[†††], geomagnetism with INTERMAGNET[‡‡‡] and others)
3. *National Research Infrastructures*, made up of Research Infrastructures (RIs) providing data and services.

The details of the EPOS architecture are out of the scope of this document and have already been discussed by Jeffery et al.[2]. However it's worth of evidencing the high density of data and stakeholders (more than 250 Research Infrastrucures, about 150 institutions, more than 120 laboratories, thousands of GPS and seismic sensors, several Petabytes of data to be managed)[§§§] creates an almost unprecedented degree of complexity in the management of data, users, software and resources in the solid Earth science domain.

4. Metadata Dimensions

As discussed by Jeffery et al. [2], in order to be able to integrate at Integrated Core Services (ICS) Level the Data, Dataproducts, Software and Services (DDSS) provided by service providers and namely by Thematic Core Services (TCS), the main ICS-C system needs an organised and homogenised view of all the elements it has to deal with (e.g. data, resources, web services endpoints etc.). The chosen approach to provide such homogenised view can be referred to as Metadata driven brokering[3]. Its main element is a metadata catalogue which contains all the metadata that enable the system to manage the user request. Four main dimensions of metadata (Fig 1b) have been identified: a) Users metadata, to profile users and describe their interaction with the system; b) Processing metadata, to enable the system to perform all processing-related operation; this includes for instance description of remote processing environment, or information about the creation of workflows; c) Data metadata, to describe the data and dataproducts and enable the system - for instance - to provide discovery functionalities for the end users; d) Resource metadata, to describe all resources the system has to interact with; it includes description of local or national research infrastructures, or 3rd party facilities. A detailed description is provided by Jeffery et al.[2].

5. Why CERIF

Since the first phase of EPOS (EPOS-Preparatory Phase Project, 2010-2014), a main system architecture was proposed. It evolved through three stages of refinement with much consultation both with the EPOS community representing EPOS users and participants in geoscience and with the overall ICT community [4].

The central role of metadata was immediately clear, and in order to manage it a three layer metadata structure was adopted[5] (Fig 1c). One main component of such a structure is the contextual layer, that in the framework of EPOS was implemented with the CERIF model. CERIF stands for "Common European Research Information Format", it is an EU recommendation to members states, and was developed with the support of the European Commission (EC). It is now maintained by euroCRIS[****]. CERIF was chosen for the following reasons: (a) its entities and attributes cover the real-world entities and attributes of interest to EPOS; (b) it has formal syntax and declared semantics; (c) it has been demonstrated to be interoperable with metadata standards from the geosciences [11,12]; (d) it has been demonstrated to generate metadata suitable for open government data [10]; (e) it is extensible (while an interoperating core is preserved); (f) it is evolving and supported by euroCRIS; (g) it is used widely with active user communities; (h) the concept of base entities and linking entities (with roles and temporal duration) permits a fully connected graph structure unlike the flat or hierarchical structures of other metadata standards - in addition the temporal timestamps allow to keep provenance information; (i) it is natively multilingual; (j) the semantic layer supports definitions of terms such as roles in relationships and standard values for attributes; (k) the semantic layer also has a structure allowing multilingual crosswalking between terms which is important for interoperability.

[†††] http://www.orfeus-eu.org/eida/
[‡‡‡] http://www.intermagnet.org/
[§§§] statistics available at the RIDE database web page (Research Infrastructure Database for EPOS) https://www.epos-ip.org/ride
[****] http://www.eurocris.org/cerif/main-features-cerif

6. Mapping EPOS elements to CERIF: Challenges

CERIF is - as discussed - a very flexible and rich model, that can represent a number of concepts in an integrated way. It needs however to be adapted in order to properly represent the main concepts of the EPOS ecosystem. One of the main challenges is therefore the mapping of EPOS concepts or metadata (categorised in the four dimensions: Users, Data, Software, Resources) to the CERIF model. If the mapping is not exhaustive, i.e. not all metadata from EPOS can be represented in CERIF, then a gap analysis and a further extension of the model is required.

6.1. CERIF mapping previous work

The work of mapping and extending, after a gap analysis, the community metadata to CERIF has already been done in other contexts.

In the case of Cerif4Datasets (C4D)[††††], the goal was to use CERIF to capture the metadata of research datasets, and integrate such metadata with that held on research projects and research outputs available on a central CERIF repository [6]. Such exercise demonstrated the possibility of mapping research datasets to CERIF, and at the same time evidenced the necessity of CERIF extensions or re-versioning to handle some attributes (e.g. spatial resolution) [7]. Some of the recommendation have been taken into account in CERIF 1.6, which can handle, for instance, the spatial coverage of a dataset through the *cfGeoBBox* entity.

Likewise, the MERIL[‡‡‡‡] project worked on the representation of Research Infrastructure (RI) concept in CERIF, in order to "create an inventory of openly accessible research infrastructures (RIs) of more-than-national relevance in Europe across all scientific domains". Documents from euroCRIS are available, which provide an overview of the RI mapping work done by MERIL [8], evidencing what entities and attributes from CERIF can be used to map a Research Infrastructure.

Building on such results and making use of their outcome, we will show how the mapping work is being set up and carried on in the EPOS context.

6.2. EPOS TCS communities heterogeneity and metadata baseline

Metadata by it's inherent nature presents different interpretations from the many standards (e.g. ISO, INSPIRE) available to many scientific communities. Due to these different requirements and interpretations, many profiles have evolved as either minor and/or major extensions of the well established metadata profiles.

This presents great complexity for establishing a common metadata catalogue for the EPOS ecosystem which seeks to bring together the resources of a diverse and extensive TCS community with many different metadata profiles implemented at different levels of granularity to the standards themselves. For example, an ISO 19115 or INSPIRE metadata implementation at different mature TCSs have different granularities due to the different interpretations of the core metadata profile and also what extensions each individual TCS has applied to the core elements. Another complexity is with the less mature TCSs who may not have a common metadata profile for their community. So, heterogeneity is at multiple levels: a) different standard metadata profiles, b) different interpretations of the same standard metadata profile and c) also no standard metadata profile in use.

All of the above creates a level of complexity, on cataloguing the required metadata, vocabularies/common semantics, data services endpoints - standard web services (i.e. to a standard pattern e.g. OGC compliant services) and also how and what metadata to expose through the EPOS ecosystem.

To enable EPOS create a level of harmonisation for these different profiles from the TCS, to communicate the minimum level of metadata required from the TCS to support the EPOS system and to facilitate its mapping to our metadata model - CERIF - we undertook the task of constructing a *metadata baseline* to provide this context.

[††††] https://cerif4datasets.wordpress.com/

[‡‡‡‡] http://portal.meril.eu/converis-esf/publicweb/startpage?lang=1

The EPOS metadata baseline includes a) the core elements of standard and common metadata profiles (ISO 19115, INSPIRE), b) elements of interest to the EPOS community that need to be represented into the metadata catalogue (e.g. Research Infrastructure) and c) elements attributed by certain TCSs for their resources (e.g. Software, Equipment) required to support the EPOS system but not provided within these established metadata profiles. The EPOS metadata baseline in the actual version contains the elements shown in Fig. 2 and has the following goals:

1. A metadata baseline that will be used to communicate with the various TCS as to the *minimum* metadata required (predominantly ISO 19115 and INSPIRE core elements with EPOS extensions)
2. A metadata baseline that defines what in a *minimum* will be mapped to CERIF
3. A metadata baseline - *minimum* - that captures core metadata elements required for populating the EPOS metadata catalogue and hence developing the EPOS system.

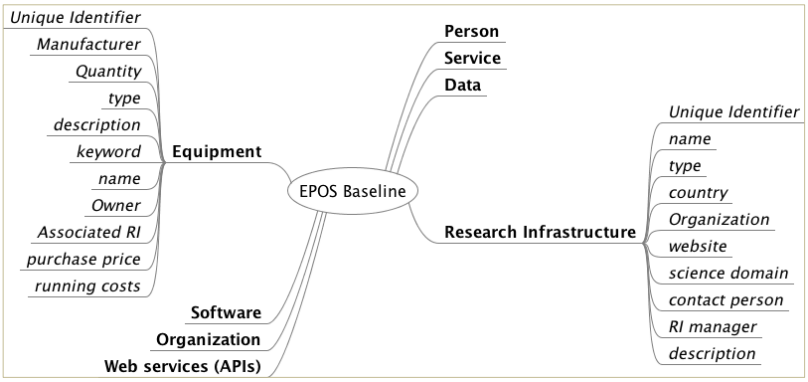

Fig2. A visual representation of the EPOS Metadata baseline. In the picture above the elements identified so far are shown in bold: *Equipment*, representing equipment from communities facilities (e.g. seismic networks, laboratories); *Software*, representing avaiable software to be integrated into ICS or discoverable for download at ICS level; *Organization*, representing an institution or natinal / international organization; *web services (APIs),* Online interfaces (APIs) enabling a user or a machine to programmatically access the given resource; *Person*, representing a human being (not legal entity); *Service*, that in compliance with CERIF model describes an offer of an RI to the community; *Data*, representing data or data products, and whose attributes are based on INSPIRE standards and suggestions by RDA[§§§§]; *Research Infrastructure*, representing the research infrastructure concept. *Equipment* and *Research Infrastructure* elements, for which we show a mapping in the present work (Table I and Table II), are shown in detail with the list of specific attributes.

6.3. Representation of the Research Infrastructure and equipment concepts

The concept of Research Infrastructure (RI) plays a central role in EPOS, therefore this entity has been chosen as starting point in our analysis. An important goal in EPOS is to leverage existing resources and metadata minimising the disruption of used standards and technologies. For this reason we decided to start with the widespread and broadly accepted INSPIRE definitions plus a few EPOS additional elements as our baseline. Boldrini et al [11] already demonstrated the feasibility of bidirectional crosswalks between INSPIRE based information infrastructures, namely ISO 19115, and CERIF and proposed extensions to both metadata models. In particular the latest version of CERIF, 1.6, requires some adjustments to describe "information for the re-use and accessibility of datasets" whereas in the INSPIRE profile of ISO19115 it is not possible to relate datasets and scientific results to projects and RIs, thus missing fundamental bricks to describe research information.

According to Dvořák et al.[13] RIs are modeled as CERIF facilities. RIs provide services and contain equipment. Furthermore, they are linked to a hosting organisation, provide a contact person, and are classified along four dimensions, namely the scientific domain, a category, a status, and a type.

[§§§§] "The Research Data Alliance (RDA) builds the social and technical bridges that enable open sharing of data", full description available at https://europe.rd-alliance.org/

Table 1 Mapping of EPOS Research Infrastructures to CERIF

Element (requirement)	**Vocabulary**	**CERIF Entity/Relationship**	**CERIF Attributes**
Identifier (m)	(not needed)	*cfFacil*	*cFacilID*
Name (m)	(not needed)	*cdFacilName*	*cfFacilID, cfName*
Type (o)	Scheme and Class to characterise the "type"	*cfClass,cfFacil_Class cfClassCheme*	*cfClassScheme, cfClass*
Postal Address (m)	Scheme: CERIF_ENTITIES Class: POSTAL_ADDRESS	*cfFacil_Paddr, cfPaddr*	*cfFacilID, cfPAddrID, cfCountryCode, cfAddrLine1, cfPostCode, cfCityTown, cfStateofCountry*
Website (o)	(not needed)	*cfFacil*	*cfFacilURI*
Description (m)	(not needed)	*cfFacilDescr*	*cfFacilID, cfDescr*
Science Domain (m)	Scheme and Class to characterise the "scientific domain"	*cfClass,CfFacil_Class, cfClassCheme*	*cfClassScheme, cfClass*
Affiliation (m)	Scheme: ORGANISATION_RESEARCH_INFRASTRUCTURE_ROLES Class: OWNER	*cfOrgUnit_Facil*	*cfOrgUnitiD, cfFacilId*
Manager (o)	Scheme: PERSON_RESEARCH_INFRASTRUCTURE_ROLES Class: MANAGER	*cfPers_Facil*	*cfPersID, cfFacilID*
Costs (o)	Scheme: RESEARCH_INFRASTRUCTURE_COSTINGS Class: OPERATING_COSTS	*cfFacil_cfFund, cfFund*	*cfFacilID, cfFundID, cfAmount, cfCurrCode*
Financial Contact Person (m)	Scheme: PERSON_RESEARCH_INFRASTRUCTURE_ROLES Class: FINANCIAL_CONTACT	*cfPers_Facil*	*cfPersID, cfFacilID*
Legal Contact Person (m)	Scheme: PERSON_RESEARCH_INFRASTRUCTURE_ROLES Class: LEGAL_CONTACT	*cfPers_Facil*	*cfPersID, cfFacilID*

We identified a set of additional attributes characterising a Research Infrastructure and performed a preliminary classification into mandatory (m) and optional (o). For each of the attributes we describe a possible mapping onto CERIF entities. Table I illustrates the Research Infrastructure entity with the list of attributes, their definitions and the corresponding CERIF mapping. The proposed mapping shows how each RI can be classified with respect to its type and scientific domain adopting the CERIF semantic layer. Also, this suggests the adoption of common Controlled Vocabularies which in the heterogeneous landscape of EPOS may facilitate interoperability. The scope of such a vocabulary, for instance describing scientific domains, would be for sure of broader interest.

Furthermore, we compiled a set of attributes that define Instruments and Equipment in the EPOS context. Equipment is characterised by its mandatory attributes, i.e. the identifier, a description, an owning organisation, an associated research infrastructure, and an equipment type. In addition, we identified a manufacturer, a purchase price, running costs, keywords and a name as potential interesting and optional information. In the CERIF model we found that our attributes are in some cases directly linked to equipment as for description, keywords, and name. In other cases our attributes had to be modeled through linking existing CERIF entities. In particular, we used the linking between *cfEquip* and *cfOrgUnit* to model the ownership of equipment and we used the linking between *cfEquip* and *cfFacil* to model a research infrastructure associated to equipment. In both cases the vocabulary for classes and class schemes published along CERIF 1.5 already contains appropriate vocabulary. Very specific to the EPOS domain are the types of equipment we expect to hold in the database. Certain types of instruments are not available in every institution because their maintenance or the devices themselves are expensive. So there is an interest to make available equipment visible and searchable. To facilitate the search we decided to associate controlled vocabulary expressing a type of equipment. In the CERIF model the class entities are used to model controlled vocabulary. Our next equipment attribute - the manufacturer - is an organisation and we use again the linking to *cfEquip* to store this information.

Unfortunately, the CERIF semantics provide at the moment no vocabulary to describe manufacturers and we think that in this case the vocabulary could be extended.

Table 2 Mapping of EPOS equipment to CERIF

Element (requirement)	**Vocabulary**	**CERIF Entity/Relationship**	**CERIF Attributes**
Identifier (m)	(not needed)	*cfEquip*	*cfEquipID*
Manufacturer (o)	Scheme: ORGANISATION_TYPES Class: MANUFACTURER	*cfEquip_cfOrgUnit, cfOrgUnit, cfOrgUnitname*	*cfEquipID, cfOrgUnit, cfName*
Description (m)	(not needed)	*cfEquipDescr*	*cfLangCode, cfDescr*
Keyword (o)	(not needed)	*cfEquipKeyw*	*cfLangCode, cfKeyw*
Name/Title (o)	(not needed)	*cfEquipName*	*cfLangCode, cfName*
Type (m)	EPOS specific	*cfClass, CfEquip_Class, cfClasscheme*	*cfClassScheme, cfClass*
Owner (m)	Scheme: ORGANISATION_RESEARCH_INFRASTRUCTURE_ROLES Class: OWNER	*cfEquip_cfOrgUnit, cfOrgUnit, cfOrgUnitname*	*cfEquipID, cfOrgUnit, cfName*
Associated Research Infrastructure (m)	Scheme: RESEARCH_INFRASTRUCTURE_RELATIONS Class: PROVISION	*cfEquip_Facil, cfFacil, cfFacilname*	*cfEquipID, cfFacilID, cfName*
Purchase Price (o)	Scheme: RESEARCH_INFRASTRUCTURE_COSTINGS Class: CONSTRUCTION_COSTS	*cfEquip_cfFund, cfFund*	*cfEquipID, cfFundID, cfAmount, cfCurrCode*
Running Costs (o)	Scheme: RESEARCH_INFRASTRUCTURE_COSTINGS Class: OPERATING_COSTS	*cfEquip_cfFund, cfFund*	*cfEquipID, cfFundID, cfAmount, cfCurrCode*

Initial and running costs of equipment can be described by linking *cfEquip* and *cfMeas* and describing this relation using the "research infrastructure costings" scheme published with CERIF 1.5. While these modeling of costs can work, we think money flows should better be kept in one entity "*cfFund*" and money associated to equipment should be modeled through the linkage of *cfEquip* and *cfFund*. We therefore suggest to open the vocabulary grouped in the class scheme "research infrastructure costings" for use in the linking relation *cfEquip_Fund*. An overview of the mapping of the equipment attributes is shown in Table II.

7. Discussion and Conclusion

This paper presents the usage of the CERIF model in the context of the EPOS e-Research Infrastructure integration plan. In order to understand the challenges posed by its complex organization, that includes thousands of stakeholders and hundreds of institutions and data providers, the EPOS architecture is shown, and the choice of CERIF is discussed. To include into CERIF all the elements relevant to EPOS (i.e. elements that enable the EPOS Integrated Core Service system to satisfy the users' requests), a metadata baseline, representing the minimum set of metadata elements to be mapped into the catalogue, is presented and discussed.

Mapping examples from the *EPOS metadata baseline* to CERIF are shown and discussed. Such mapping exercise raised a number of issues, questions and suggestions.

First, while mapping the Research Infrastructure element, the need of a classification to define the type and scientific domain of a RI emerged. We suppose this necessity in the EPOS landscape might be of broader interest. In

addition, we encountered issues in describing the role of the Financial and Legal contact person. There is a need to characterise these roles with the two new classes FINANCIAL_CONTACT and LEGAL_CONTACT in the scheme PERSON_RESEARCH_INFRASTRUCTURE_ROLES.

A second issue was raised, while mapping the Equipment and Research Infrastructure elements: the need for a vocabulary to characterise a manufacturer of equipment. A manufacturer is an organisation - we therefore suggest to create a new class MANUFACTURER in the scheme ORGANISATION_TYPES to allow classification of the relation between organisation and facility. In addition, there is a need to develop vocabulary to describe different types of equipment and we believe this task is very EPOS specific. However, we need to distinguish between "mobile" and "static" equipment and this difference is supposed to exist in other communities.

Third, although CERIF supports provenance natively because of the timestamped linking entities, the integration of causual-effects relationships among the entities and activities involved and re-used across processing tasks needs to be further develped. The use of PROV-O[*****] has been trialed in the VERCE project and has also been demonstrated in biosciences. Since it is possible to map CERIF to RDF/XML and hence PROV (and such work is current in project VRE4EIC) we shall utilise both approaches in parallel before deciding which (or even both) to use. This leads towards the interoperable understanding of the scientific results and their long term preservation, in the context of the contributing artifacts, workflows and computing facilities.

Fourth, the mapping exercise shown in this work evidenced a twofold need: on one side, the need of standard tools to do the mapping; on the other side the necessity of a centralized catalogue where information about previous mappings could be obtained.

As for the tools for mapping, a very common workflow is to start with some spreadhseet based software and then to switch to formal languages that map from a standard format (e.g. XML) to another format. An example language is XSLT. A software tool that would enable a developer to do the mapping in a graphical or tabular way, express the mapping in more than one formal language, and then apply the mapping with the possibility of choosing among several input and output formats (including Linked Open Data standards), would incredibly facilitate the mapping activity and make the mapping ready to be shared with other stakeholders. Some example exist, as in the case of 3M Mapping tool by FORTH[†††††], that hoewever needs further developments in authors' opinion to permit mapping from/to several input and output format.

As for the "catalogue of mappings", a serious issue faced while doing the mapping was to search for previous work done on the elements to be mapped in the EPOS context, as in the case of the *Data* element in the *EPOS metadata baseline*, which is based on the ISO 19115 and INSPIRE standards. Such mapping was already done to some extent in the cited papers[11], but detailed information were not available or simple to find. A centralized "canonical" archive, for which euroCRIS seems to be the evident maintainer, where mappings are made available in a standard format and where each mapping is linked to publications dealing with it, would help both CERIF users, when doing the mapping work, and CERIF maintainers (e.g. euroCRIS community), to have a clear vision of the diffusion and work done on CERIF in a pan-European dimension.

Finally and similarly to the previous issue, we discovered that CERIF related software developed by the many initiatives at European level is hard to find in some cases. Software is scattered across different public repositories like github, google code, and others and there is no central place to go to search for software. A catalogue of open source software available from the euroCRIS website, would facilitate the development of new software and foster the reuse of work already done. To make CERIF related software discoverable another option would be to encourage authors to publish their program code through DataCite and cite their work in publications.

[*****] https://www.w3.org/TR/prov-o/

[†††††] http://www.researchspace.org/home/mapping

Acknowledgements

The authors acknowledge the work of their colleagues in EPOS working packages ICS-TCS integration and interoperability (WP6) and ICS development (WP7), in particular the work of Roberto de Virgilio (INGV) and of the BGS. The team has received much support from euroCRIS and particularly from Valerie Brasse. Funding for this work was provided through the EU Horizon 2020 project “EPOS-IP”. Finally, we want to acknowledge the team of VRE4EIC project.

References

1. Bailo, D.; Jeffery, K.G.; Spinuso, A.; Fiameni, G.,Interoperability Oriented Architecture: The Approach of EPOS for Solid Earth e-Infrastructures, e-Science (e-Science), 2015 IEEE 11th International Conference on , vol., no., pp.529-534, Aug. 31 2015-Sept. 4 2015 DOI: 10.1109/eScience.2015.22
2. Jeffery, Keith G., and Daniele Bailo. "EPOS: Using Metadata in Geoscience." Metadata and Semantics Research. Springer International Publishing, 2014. 170-184 DOI: 10.1007/978-3-319-13674-5_17
3. S.Nativi, K.J. Jeffery, R.Kostela, RDA: Brokering with Metadata https://www.researchgate.net/publication/271646528_RDA_Brokering_with_Metadata
4. Jörn Lauterjung et al.. (2015). Report on EPOS e-infrastructure prototype. Zenodo. DOI: 10.5281/zenodo.19175
5. Jeffery, K., Asserson, A., Houssos, N., & Jörg, B. (2013). “A 3-Layer Model for Metadata.” Proc. Int’l Conf. on Dublin Core and Metadata Applications 2013, 3–5.
6. Bokma, Albert , Garfield, Sheila , Nelson, David , Omran, Esraa , Corcho, Oscar , “Cerif4Datasets (C4D) – Utilising Semantics for the Discovery and Exploration of Datasets in Research”, CRIS2012: 11th International Conference on Current Research Information Systems (Prague, June 6-9, 2012), euroCRIS, http://hdl.handle.net/11366/94
7. Brander, S., Clements, A., McCutcheon, V., Cranner, P., Henderson, R., & Ginty, K. (2013). CERIF for Datasets (C4D)-Linking and contextualising publications and datasets, and much more. DOI: 10.1007/978-3-319-08425-1_11
8. Brasse, Valérie , MERIL: An e-infrastructure to connect Research Infrastructures, 10th euroCRIS Strategic Seminar: “Horizon 2020 and Beyond” (Brussels, Sep 10-11, 2012), http://hdl.handle.net/11366/297
9. Beckers, Paul , Jägerhorn, Martin , Höllrigl, Thorsten “Advances in Sharing and Managing Knowledge about European Research Infrastructures”, RIS2012: 11th International Conference on Current Research Information Systems (Prague, June 6-9, 2012), http://hdl.handle.net/11366/91
10. Jeffery, K., Asserson, A., Houssos, N., Brasse, V., & Jörg, B. (2014). From Open Data to Data-Intensive Science through CERIF, *00*. DOI: 10.1016/j.procs.2014.06.032
11. Boldrini, E., Luzi, D., Nativi, S., & Pecoraro, F. (2014). Integrating CERIF entities in a multidisciplinary e-infrastructure for environmental research data. *Procedia - Procedia Computer Science*, *33*, 183–190. http://doi.org/10.1016/j.procs.2014.06.031
12. Ginty, Kevin; Kerridge, Simon; Fairley, Paul; Henderson, Ryan; Cranner, Paul; Bokma, Albert; Garfield, Sheila: CERIF for Datasets (C4D) – An Overview. In: Jeffery, Keith G; Dvořák, Jan (eds.): E-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production: Proceedings of the 11th International Conference on Current Research Information Systems (June 6-9, 2012, Prague, Czech Republic). Pp. 53-60. ISBN 978-80-86742-33-5. http://www.eurocris.org/Uploads/Web%20pages/CRIS%202012%20-%20Prague/CRIS2012_5_full_paper.pdf
13. Dvořák, J (2013): Research Information: The euroCRIS Context; Presentation at the "Use Science: Open Infrastructure to Foster Collaboration between Industry and Academia", 28-20 November 2013, http://hdl.handle.net/11366/41
14. Europe's environment, “An assessment of assessments”, 2011, EEA — European Environment Agency, available at http://www.eea.europa.eu/publications/europes-environment-aoa/#parent-fieldname-title