Establishing Core Concepts for Information-Powered Collaborations

Establishing Core Concepts for Information-Powered Collaborations

Luca Trani[a,1,2,*], Malcolm Atkinson[b,2], Daniele Bailo[c,3], Rossana Paciello[c,3], Rosa Filgueira[d,5,4]

[a] *Utrechtseweg 297, 3731 GA, De Bilt, The Netherlands*
[b] *Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, UK*
[c] *Via di Vigna Murata 605, 00143, Roma, Italy*
[d] *EPCC, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, UK*

Abstract

Science benefits tremendously from mutual exchanges of information and pooling of effort and resources. The combination of different skills and diverse knowledge is a powerful capacity, source of new intuitions and creative insights. Therefore multidisciplinary approaches can be a great opportunity to explore novel scientific horizons. Collaboration is not only an opportunity, it is essential when tackling today's global challenges by exploiting our fast growing wealth of data. In this paper we introduce the concept of Information-Powered Collaborations (IPC) – an abstraction that captures those requirements and opportunities. We propose a conceptual framework that partitions the inherent complexity of such dynamic environments and offers concrete tools and methods to thrive in the data revolution era. Such a framework promotes and enables information sharing from multiple heterogeneous sources that are independently managed. We present the results of assessing our approach as an IPC for solid-Earth sciences: the European Plate Observing System (EPOS).

Keywords: information and knowledge exchange, semantic interoperability, multidisciplinary collaborations, standard vocabularies, DCAT

1. Introduction

Cooperation and collaboration have characterised the organisation of work in various contexts throughout history. Consequently, the support for collaborative work has been

*Corresponding author
Email address: trani@knmi.nl (Luca Trani)
[1] Department of R&D Seismology and Acoustics, Royal Netherlands Meteorological Institute (KNMI)
[2] School of Informatics, University of Edinburgh
[3] Istituto Nazionale di Geofisica e Vulcanologia, Rome
[4] British Geological Survey, Edinburgh
[5] EPCC, University of Edinburgh

investigated for a long time by scientific disciplines such as the Computer Supported Co-operative Work (CSCW). Since the mid 80s a rich CSCW literature produced several theories and approaches proposed to model and improve collaborative work sustaining sharing of knowledge and expertise [1, 2, 3]. The importance of scientific collaborations is not only well-recognised but it is encouraged and fostered, *e.g.* by policy makers and funding bodies, as a way to improve impact, to achieve cost-efficiency and to tackle the pressing data challenges faced by nearly all scientific disciplines. Collaborations based on information sharing, contribute different viewpoints and combine skills and intellectual efforts to tackle the increasing complexity of contemporary scientific challenges. We propose a conceptual framework that combines two ingredients – collaborations and data – to help establish *Core Concepts* (CC) underpinning *Information-Powered Collaborations* (IPC). Cooperation among diverse actors carries inherent socio-technical issues and requires us to maintain 'a common terminology and shared knowledge base' that enable communication and understanding [4]. The framework proposed in this paper provides a set of tools to build and maintain a common vocabulary and a shared information space.

Data and research collaborations are strongly interrelated, and collaboration starts with sharing – data sharing has received considerable attention in the last decade being widely recognised as an accelerating factor for the scientific progress [5]. Nevertheless, data sharing is just one aspect underpinning research collaborations. Equally important are: sharing of methods, context and best practices; understanding of implicit communication rules, norms and prior knowledge that form the culture of the involved scientific communities (*Designated Communities*) [6]. Many aspects of the culture, such as formalised methods and data-access rules may be represented as shareable data so that extensive distributed collaboration can be better supported.

Building research collaborations is a major endeavour that requires time and investments that increase rapidly with the diversity and the number of involved parties. Retaining the value of those investments, sustaining and maintaining efforts over time are necessary strategic choices. The management of research collaborations ought to interface and account for the organisational structures present in each community. Different strategies may be needed to address these issues.

Our framework builds on autonomous sources of information and a set of formal agreements to support IPC, combining intellectual effort and pooling resources and expertise from multiple independent organisations. The framework is based on: a) a *Canonical Core* that holds b) *Core Concepts* and connects to c) a set of dynamic *Boundary Regions* needed to sustain a community's agility and innovative drive.

Such a framework enables holistic views of the autonomous sources whilst preserving

specialised domain specific views. In this paper we present such a framework and describe its application in the context of solid-Earth sciences. The remainder of the paper is organised as follows: in Section 2 we present the rationale that motivated this effort; Section 3 contains related work; in Section 4 we describe our conceptual framework; in Section 5 we introduce a research infrastructure for solid-Earth sciences, the European Plate Observing System (EPOS), and illustrate an application of our methodology in that context; Section 6 provides a preliminary evaluation of the presented approach; finally in Section 7 we draw conclusions and outline future work.

2. Motivation

2.1. Information-Powered Collaborations

We define the concept of **Information-Powered Collaborations** (IPC) below.

Definition 1. *Information-Powered Collaborations (IPC): are complex, dynamic and heterogeneous environments that enable information sharing among actors (e.g. researchers, scientists, practitioners) from independently managed organisations (e.g. research institutes, resource providers), thereby supporting knowledge and expertise exchange in a multidisciplinary context. The resulting collective knowledge can be harnessed to accomplish common scientific goals and foster novel scientific viewpoints drawing on the integration of the diverse perspectives.*

IPC is an abstraction that represents a typical modern research context characterised by rich interactions, exchanges and complex dynamics. Traditionally the research scene was dominated by research groups in controlled environments, with limited interactions with their peers [7]. The data revolution has deeply impacted every domain demanding a paradigm shift where collaboration is essential to manage the amount of data and to interpret the derived information. The IPC can offer a means to address and tackle today's challenges stimulating and facilitating *pooling of knowledge* (as well as data and information). To achieve this, the IPC must fulfil a number of requirements, as we illustrate below.

2.2. Use cases

In this section we present a selection of use cases and the corresponding requirements that an IPC should fulfil.

2.2.1. Resource discovery

Resource discovery – implies the search of high level descriptions (metadata) carrying information for instance, about type, name and origin of a resource. It entails operations such as selection and filtering matching specified criteria. Examples of a multi-faceted search crossing domains: `FIND` all time series catalogued since `date, time` giving geochemical emission, seismic activity and surface movement for Etna; or `FIND` the seismic events in 2017 in Southern Europe together with geology, Global Navigation Satellite System (GNSS) velocity and satellite data correlated with those events.

2.2.2. Resource evaluation

Resource evaluation – requires deeper descriptions of resources (*e.g.* domain-specific and contextual metadata) [8, 9]. It exploits additional metadata fields beyond the classification of a resource in order to query, select, filter actual instances of resources according to desired characteristics. Example: `FIND` all the seismic events with magnitude $M > 5$, that occurred in a time-window `(Tw)`, in a specific region `(Re) AND` the related primary data (seismic waveforms) with fewer gaps than 5% in `Tw AND` the GPS displacement maps associated with `(Tw, Re)`.

2.2.3. Scientific methods

Scientific methods support – helps collaborating teams of experts create and refine methods that draw on the diverse resources and data collections. It promotes the formalisation and automation of these methods, typically as scientific workflows [10], while supporting critical procedures to deliver good quality evidence contributing to the shared knowledge.

Example: develop methods and models to reveal the impact on seismic hazard from mineral extraction methods. The authoring system consults the metadata catalogue to help the method developer make choices, detect defects and plan enactment. The enactment system consults the metadata catalogue to verify compliance with policies, to plan the optimal deployment and annotate provenance records. The provenance system links with the catalogue, mainly via identifiers, to support diagnostics, validation, reproducibility and evidence qualification.

2.3. Supporting shared agreements

Metadata catalogues play a central role in our framework – the currently agreed set of instances of Core Concepts is represented by these catalogues. Standard vocabularies provide a vital element of the Core Concepts. It is widely recognised that they help fulfilling requirements 2.2.1 and 2.2.2. In particular, vocabulary profiles enable validation via lists

of allowed values, cardinality of elements and specifying detailed application contexts. However, in order to achieve semantic interoperability and enactment of workflows 2.2.3 exploiting cross-domain resources, the mode of employment of such vocabularies must also be specified. This requires the definition of agreements about the interpretations and meanings of the values associated with vocabulary terms, and the formalisation of such agreements in the shared vocabulary. The latter can be achieved by introducing formal restrictions and constraints expressed for instance in OWL[6], SHACL[7] or SHEX[8].

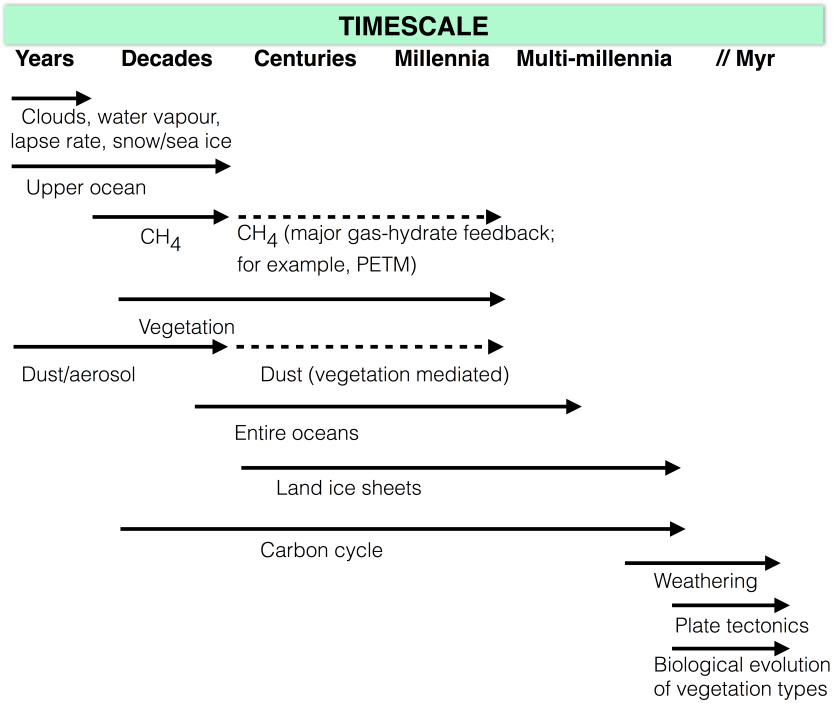
Achieving shared agreements on the interpretation of vocabulary terms in multidisciplinary environments is not a trivial task. Even a common concept such as `time` can carry diverse semantics depending on the temporal reference or the calendar used in specific context. For instance, in archeology or geology time is often expressed counting years backwards from a reference date. In a lunisolar calendar (*e.g.* Chinese Calendar) time is expressed according to astronomical phenomena. Those reasons inspired domain-specific formalisations, *e.g.* for geological timescales [11], and extensions in conventional representations such as OWL-Time[9] to include non-Gregorian calendars [12]. Figure 1 provides an example of the diversity in time scales which are present in solid-Earth sciences. Each of those might be associated with different reference systems and therefore a different semantics of time. The deployed instruments may resolve time with sub-microsecond resolution to triangulate signal sources. A conceptual framework from geological, through historical to observational time needs clarity about the transitions and correspondences.

In order to support multi-disciplinary, multi-organisational and multi-national collaboration the underlying concepts must be recognised and agreed. These are often formalised as ontologies [14]. Collaborative development of such ontologies often reveals variations and encourages refinement of such concepts, illustrating the kind and scale of investment needed to build, agree and adopt the Core Concepts.

3. Related work

We observe initiatives that aim at supporting multidisciplinary research collaborations and investigate relevant enabling technologies and standards. We review those initiatives and technologies in terms of their contribution to the three aspects of support: 1. agreeing a common set of concepts; 2. representing that agreement in human readable and machine actionable forms; and 3. provisioning of tools and platforms for collaboration. In most of

[6] `www.w3.org/owl/`
[7] `www.w3.org/TR/shacl`
[8] `https://shexspec.github.io/spec`
[9] `www.w3.org/TR/owl-time/`



Source: [13]

Figure 1: Diversity in time scale nomenclature and precision experienced in research that engages with the distant past as well as the present, such as the solid-Earth sciences. This is just part of the range encountered by sciences that observe to sub-microsecond resolution for today's observations to resolve hypothesised models spanning billions of years.

the cases these are intricately bound together but we show the value of considering them separately.

3.1. Agreeing common concepts

This predominantly manifests as defining agreed vocabularies and importing, extending or merging existing ones.

3.1.1. Schema.org

Schema.org is a vocabulary created in 2011 by Google, Microsoft and Yahoo to describe Web resources and improve the search of content on the Web, thus assisting search engines as they interpret pages in different contexts. Since its conception Schema.org has grown into a popular mechanism to represent structured data on the Web; it is supported by many tools and includes a variety of domains [15]. Schema.org is constituted by a hierarchy

of classes and relationships – it is compliant with RDF and reuses existing standard vocabularies such as Dublin Core. Typically it is embedded in HTML pages using Microdata[10], JSON-LD[11] and RDFa[12].

3.1.2. W3C – DCAT and DCAT profiles

W3C has invested significant effort steering the development of a vocabulary to facilitate the interoperability of catalogues published on the Web, namely the Data Catalog Vocabulary (DCAT)[13]. At present DCAT is a W3C Recommendation that has been endorsed by many players including scientific communities, policy makers and other stakeholders [16, 17]. *"By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites. Aggregated DCAT metadata can serve as a manifest file to facilitate digital preservation"*. Several profiles of DCAT have been produced to address different requirements and there is an active community supporting the uptake of their data model. Furthermore, DCAT is natively supported by catalogue platforms such as CKAN [18]. Examples of such profiles include: DCAT-AP [19] used to describe public sector datasets in Europe; GeoDCAT-AP [20] – a DCAT-AP profile describing geospatial datasets, dataset series, and services; and StatDCAT-AP – a DCAT-AP profile for statistical datasets [21]. One of the key features of DCAT is that it incorporates terms from existing and widely used vocabularies such as Dublin Core, SKOS and FOAF. This aspect increases its dissemination and facilitates adoption and uptake into existing systems.

Application profiles try to fill the gaps in the base DCAT standard. Some gaps have been identified and discussed at the "Smart Descriptions & Smarter Vocabularies (SDSVoc)" [14] workshop organised by W3C and the VRE4EIC project[15]. Although the current DCAT recommendation is recognised as a powerful tool to improve interoperability of datasets, further work and guidance are needed to extend its adoption and to tailor it to meet community requirements for particular IPC. The W3C Data Exchange Working Group (DXWG)[16] has been recently set up to collect and address requirements from the communities and help improve the DCAT data model.

[10] www.w3.org/TR/microdata/
[11] www.w3.org/TR/json-ld/
[12] www.w3.org/TR/xhtml-rdfa-primer/
[13] www.w3.org/TR/vocab-dcat
[14] www.w3.org/2016/11/sdsvoc/
[15] www.vre4eic.eu
[16] www.w3.org/2017/dxwg/charter

3.2. Representing conceptual agreements

This depends on underpinning metadata and ontologies organised, combined, referenced and analysed using formalised models.

3.2.1. Metadata and interoperability

Metadata approaches have been widely discussed as methods to enable interoperability [22, 23, 24, 25]. In the context of digital libraries the metadata interoperability issue has been recognised for a long time. As the mission of digital libraries is to acquire, preserve and provide access to a variety of heterogeneous digital objects, librarians quickly encountered issues related to the appropriate description of digital objects and developed standards and methods for their categorisation. For instance, standards-based metadata, metadata cross-walks or mappings, application profiles and metadata registries have been demonstrated to be valuable methods to enable schema-level metadata interoperability [26]. Those methods build on a classical interpretation of information organisation systems, mainly hierarchical and authoritative, thus reflecting an objectivist philosophical perspective [25]. However, that perspective has been considered inadequate to organise complex information [27]. The advent of social media stimulated collaborative approaches to metadata which exploit social tagging and yield folksonomies [28]. Such approaches reflect a social constructivist perspective of the world, they take into account heterogeneous viewpoints, fluidity of interpretation and knowledge sharing [25]. Although authoritative and collaborative, or in other words top-down and bottom-up, approaches might seem antithetic, they can coexist providing complementary perspectives and, as advocated by Gruber, lead to ontologies for folksonomy [29]. Whilst top-down approaches contribute a "'simplified" canonical view according to paradigms of classifications that have been known to humans for a long time, folksonomies recognise the existence of different possible interpretations and account for specialisations and extensions known and understood by subgroups and individuals. The Web Annotation Vocabulary[17] is an example of an ontology supporting such a collaborative approach.

Semantic interoperability entails information sharing and exchange based on negotiated meanings and expressions [22], it goes beyond the schema-level specifying how metadata records or *content values* are exchanged and used. Therefore, semantic interoperability deals with structure and includes interpretation leading to mutual understanding of concepts, relationships and their values. Alemu *et al.* argue that in order to achieve semantic interoperability metadata objects ought to be enriched with knowledge coming from collab-

[17] `www.w3.org/TR/annotation-vocab`

orative and user-driven approaches [25]. Semantic web technologies can provide the appropriate support to achieve semantic interoperability and harmonisation [24]. This depends on leveraging declared vocabularies and mechanisms to extend them; unique identifiers that help avoiding naming conflicts and duplications and the ability to express relationships among resources and elements.

3.2.2. Shapes Constraint Language (SHACL)

The Shapes Constraint Language is a recent W3C Recommendation that is rapidly gaining interest in the semantic community. Semantic languages, such as OWL [30], offer a powerful means to describe terms and how they can be used but they lack a mechanism to record the applications of such terms. The latter is particularly useful to share and reuse knowledge among communities. *Data shapes* expressed in SHACL fulfil this requirement providing an effective and flexible tool for data integration. Shapes are RDF expressions that explain how data is organised. Those expressions include allowed rules, values, patterns and offer a powerful mechanism to formalise constraints and validate data structures. They can be used as templates to model and query data structures. A number of use cases[18] for the application of SHACL are currently under discussion. The "Open Content Model"[19] (OCM) is an application context of particular interest for us. For instance, according to the OCM multiple independent applications might agree to share the same representation for common data items and allow the presence of undefined data items to account for specialisations in the diverse applications.

3.3. Platforms for collaboration

We include organisational and technical approaches to populate the conceptual space with concrete instances and examples of multidisciplinary infrastructures.

3.3.1. Computer Supported Cooperative Work

CSCW investigated the social aspects of knowledge sharing and the systems to support it. Such investigations yielded approaches to define and maintain 'common information spaces', to represent knowledge for instance by adopting a 'repository model' and/or exchange it via knowledge artifacts and 'boundary objects' [1, 31, 2, 3]. An important branch of CSCW research focused on providing access to and exchanging expertise, recognising the importance of communication and helping establish communications among 'knowledgeable actors'. For these reasons CSCW research provided a fertile ground for a number

[18] `www.w3.org/TR/shacl-ucr/`

[19] `https://www.w3.org/2014/data-shapes/wiki/Open_Content_Model_Example`

of technical solutions currently adopted in knowledge management and collaborative systems.

3.3.2. Research Data Alliance

The Research Data Alliance[20] is an international, multidisciplinary, community-driven organisation that is very active in the area of data sharing and exchange, data interoperability and data-driven innovation. Recommendations, infrastructure design, policies and various initiatives are emerging to lower the barriers to data sharing and accelerate innovation. Some of these initiatives have recently been endorsed by the European Commission[21] who recognises their importance for referencing in public procurement, in particular: 1. 'RDA Data Foundation and Terminology Model'; 2. 'RDA PID Information Types API — Persistent Identifier Type Registry'; 3. 'RDA Data Type Registries Model'; and 4. 'RDA Practical Policies recommendations'. The RDA Data Fabric Interest Group introduced the concept of Global Digital Object Cloud (DOC)[22] a virtualisation layer that exploits the components presented above to offer an architecture based on the principles of the Digital Object Architecture and fully compliant with the FAIR principles [32].

3.3.3. Virtual Research Environments (VREs) and related frameworks

Virtual Research Environments are well-known, powerful frameworks that enable collaborative science. VREs provide scientists and practitioners of communities of practice [33] with tools and working environments (or laboratories), usually accessible via the Web, that encompass data, services and computing enabled features such as processing, visualisation, communication, data access and workspaces. Such environments can be deployed in different contexts thereby serving the needs of a variety of communities, however they usually target single disciplines or closely related topics. Recent developments demonstrated the feasibility of aggregating cross-cutting resources to offer VREs as a Service in order to maximise the adoption and productivity in multidisciplinary contexts [34]. Similarly, Virtual Laboratories(VLs), Science Gateways (SGs), Virtual Organisations (VOs) and Digital Libraries (DLs) provide the necessary tools and interoperability to enable interactions and foster seamless access, usage and sharing of resources across diverse stakeholders [35, 36]. There is a substantial interest in the scientific community in VREs (VLs, SGs, VOs and DLs) that yields a flourishing scientific literature and many initiatives and research

[20] www.rd-alliance.org

[21] http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32017D1358&from=EN

[22] https://www.rd-alliance.org/group/data-fabric-ig/wiki/global-digital-object-cloud

projects. However, as shown in a recent discussion at the RDA VRE-IG[23], the terminology and the definitions, although often overlapping, are still disputed and often subject to different interpretations. In our analysis, whilst acknowledging the diverse flavours, we use those terms interchangeably. Such systems deal with the human-computer interactions and socio-organisational issues as well as authorisation and resource management. In this paper, we assume such a context and focus on supporting the process needed to build an underpinning alignment of concepts and information.

3.3.4. Virtual Observatories (VOs)

The concept of Virtual Observatory was first introduced by the astronomers as a means to discover, access and process data seamlessly [37, 38]. The goal was to provide an abstraction layer on top of astronomical data provided by independent organisations following the analogy of the World Wide Web. The astronomy community produced a predominant example of successful, long-term collaboration led by the International Virtual Observatory Alliance (IVOA). IVOA discusses and promotes standards for interoperability, protocols for data access and exchange. Since its establishment in 2002 it has supported the astronomy community to establish innovative technical solutions at global scale, disseminate results and promote effective collaborative working practices [39].

Other examples of VO are: the CLARIN Virtual Language Observatory[24] targeting language resources [40]; and the Web Observatory (WO) – a large system that enables multidisciplinary Web Science. WO focuses on data *about the Web* to study and understand the evolution of the Web anticipating future trends and developments [41, 42]. The Web Observatory provides data analytics tools that exploit, integrate and harmonise data originating from heterogeneous sources. WO is a clear example of distributed environment for collaboration and knowledge sharing that for its scale and scope could not be managed as a centralised data warehouse. An important feature of the Web Observatory is that it provides access to and preserves existing repositories. Hence, WO can be considered as an IPC with the related challenges. Tiropanis *et al.* [43] report their experience designing and building an architecture for a network of Web Observatories at the University of Southampton (SUWO). They recognise several challenges, for instance related to data ownership and access control. Wang *et al.* [44] propose a data cataloguing framework, called WDFed, that leverages Linked Data and RESTful APIs to enable discovery and use of heterogeneous data on the Web. This framework is implemented in the WO at SUWO and exploits

[23] https://www.rd-alliance.org/group/virtual-research-environment-ig-vre-ig/post/looking-authoritative-definitions-vre-vlab-science

[24] https://vlo.clarin.eu/

the DCAT data model to harmonise and represent collections of multidisciplinary datasets.

3.3.5. The Global Earth Observation System of Systems (GEOSS)

GEOSS is a global initiative coordinated by the Group on Earth Observation (GEO) to build a large-scale network of content providers into a single overarching system. It embraces the most important existing infrastructures for Earth Observation at a global scale. GEOSS adopts the well-known System of Systems (SoS) approach where many autonomous, independent systems are coherently networked and co-operate to achieve common goals [45]. The GEOSS Common Infrastructure (GCI) is the e-infrastructure that underpins GEOSS and leverages the distributed independent resources, harmonising data and models, providing access to resources, applications and products. The GCI exploits a brokering approach to provide users with transparent access to the distributed resources [46]. The concept of SoS captures the common issue of integrating many independent, autonomous systems in order to achieve a global common goal. GEOSS aims to provide decision support tools and what-if type of analysis, with information and knowledge delivery as a goal. Santoro *et al.* [47] introduce the Model Web framework that captures business processes as workflows. To address the Science-to-IT barrier issue they leverage models, workflows, vocabularies and knowledge bases. Their focus is primarily on how to combine and use those resources, whereas our focus is on how to support their construction and harmonisation leveraging Core Concepts for collaborations.

3.4. Summary of related work

We presented three aspects to support collaborations and achieve interoperability. In section 5 we will show how we leveraged existing vocabularies and formalised models such as Schema.org, DCAT, RDF and SHACL, to build a representation of the Canonical Core for EPOS. Concerning the platforms and frameworks supporting collaboration, VREs offer concrete solutions to address some of their requirements. However, in our view there are aspects that still ought to be addressed in the conceptual space underpinning such collaborative environments. The definition of IPC presented in this paper is an attempt to fill those gaps. It is inspired and builds on results deriving from the CSCW research. Although the concepts of VRE and IPC might overlap, there are characterising features that differentiate them. IPC are inherently targeting diverse disciplines and have semantic interoperability as their primary goal. IPC facilitate and promote agreements about concepts and their interpretations, thus they help communities develop a shared information space underpinning collaborations. IPC target a conceptual level that sustains communication and informed decisions between organisations, teams and individuals. Whereas VREs target an application level where the focus is on tools that should use and promote the definitions of

the conceptual space for the individual practitioner. IPC can be seen as an abstraction that helps establish collaborative behaviour whereas a VRE provides the resulting integrated environment and tools once such collaborations have been established. We argue that a VRE might increase the quality of the collaborative experience delivered to its users by leveraging the information space developed by an IPC.

Similarly, we reported approaches exploiting the SoS model whose objectives overlap with the ones of IPC. However, the primary focus of IPC is on the definition of a coherent, shared information space with the aim to foster pooling of knowledge. This poses additional constraints on the implementation. For instance, it could exclude a pure brokering approach, while remaining flexible and adaptable to existing systems. The level of information exchange characteristic of IPC requires deeper integration and agreements that cannot be delegated to an externally plugged component, but need to be negotiated with the participating parties.

4. Approach and methodology

From our assessment it appears clear that the construction of the conceptual framework that enables effective collaboration has to be led by humans. Scientific communities, users and stakeholders of an IPC assume a central role in guiding the construction and maintenance processes. Those shaping the IPC develop and maintain its conceptual core by assessing which concepts can be consistently used and interpreted across the consortium. They often proceed by importing large established vocabularies with their corresponding definitions and relationships. They need to manage the relationships between such *conceptual bundles* eventually extending or pruning them in order to meet the requirements of their IPC. They must recognise where creative diversity exists and leave opportunity for agile innovation in these conceptual spaces.

Our approach combines top-down and bottom-up strategies to formulate the agreed core set of shared concepts and achieve *semantic interoperability* in IPC. We propose that this progresses by building a *Canonical Core* (CC) that includes sufficient *Core Concepts* that are agreed and adopted to enable the principal interdisciplinary collaborations to proceed. The extensions needed beyond this CC to support innovation, experiment and local specialisations are supported by dependable relationships with the CC. Approaches based on reference ontologies have been profitably applied in more controlled contexts *e.g.* in the industry [48, 49, 50]. We build on those results to devise a solution for the challenging IPC context. The whole process exploits *co-design* bringing together data and metadata-modelling experts with domain scientists. Similarly to data models and their representations, the rules of engagement or "contracts" to participate in the IPC are critical. Such rules are discussed

and defined with the designated communities and leverage existing community standards and practices.

Figure 2 illustrates an overview of the proposed framework where the Canonical Core is a central component. As stated by the European Commission in its communication on Open Data of December 2011 [51]: "*[...]the availability of the information in a machine-readable format as well as a thin layer of commonly agreed metadata could facilitate data cross-reference and interoperability and therefore considerably enhance its value for reuse*". In our proposed framework, the Canonical Core captures agreed concepts as machine-readable metadata. The size of the core should account for several factors. It must span a sufficient range of concepts and viewpoints to meet the understood requirements for composing data, information, knowledge and methods. It must offer hooks whereby its capacity may be extended on a local experimental or specialisation basis and recipes or paths to easily incorporate successful extensions. Likewise, the core requires parsimony and consistency to make it comprehensible and manageable.

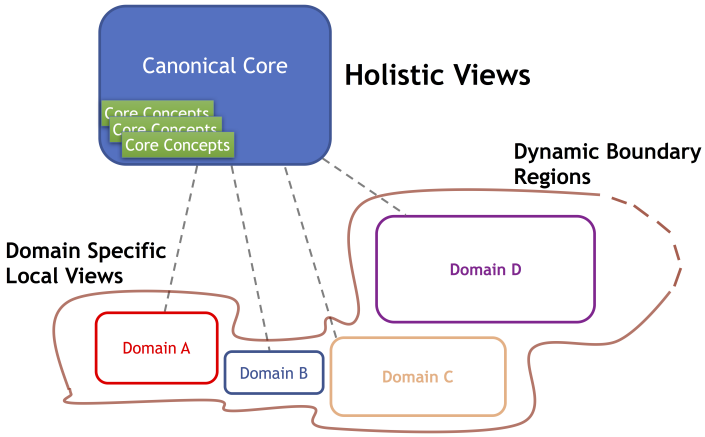


Figure 2: Overview of the framework facilitating the composition of diverse resources and their presentation to users as a coherent holistic environment. The CC provides a stable, agreed and adopted set of concepts and their relationships. The need to support innovation and handle details that are not completely adopted is met by recognising external zones of defined information models. Dynamic Boundary Regions delimit the CC from the community-specific extensions.

4.1. Dimensions of the Canonical Core

The CC represents the universe of discourse that *Designated Communities* adopt to communicate, understand and enable 'actionable' information sharing. Actionable in the

sense that it can deliver knowledge which can be understood and trusted by practitioners and interpreted by formalised automated methods. The CC is characterised by three dimensions:

1. *Conceptual definition* – what goes inside the shared information space, including the Core Concepts and their relationships.

2. *Representation* – how those concepts are represented, for instance according to a specified data model, *e.g.* using DCAT.

3. *Population* – how the CC is constructed, ingested and maintained with selected instances of the concepts therein represented and how instances are chosen among the available ones (it specifies selection criteria as not everything needs to be included at the highest level of detail).

The conceptual definition (1) constitutes an unbounded conceptual space independent from the other dimensions. In this paper we provide an approach to manage the complexity of that space, and apply such an approach to the concrete context of EPOS. We also propose a representation (2) fitting the designed space and meeting the requirements of the identified designated communities. Finally, we validate the chosen representation populated with a selection of real instances (3).

4.2. Principles underlying the conceptual definition of the Canonical Core

The conceptual definition of the CC needs to address three aims italicised below. The following principles shape the concepts, relationships and structure of the CC.

1. Achieve sufficient coverage of the behaviours required across the designated communities that the CC supports their interactions with the shared information and with each other, thereby facilitating collaboration leading to *adoption and reuse*.

2. Establish agreed interpretations of the Core Concepts that are adopted by the designated communities – when such agreements cannot be reached allocate the concepts to an extension for the relevant subcommunity coupled to the core via identified conceptual hooks – thereby achieving *harmonisation* without inhibiting innovation.

3. Validate the CC against a broad and representative set of use cases, thereby ensuring priority collaborative behaviours are enabled and achieving *trustworthiness* and *completeness*.

The volume and complexity is controlled by limiting the core to accepted and agreed material. Contenders for inclusion develop in the dynamically connected boundary regions.

The set of use cases is extended to fulfill all critical requirements and to ensure that the CC covers the essentials.

According to the principle (1), rather than building from scratch we select and import existing conceptual bundles, information spaces, boundary objects and knowledge artifacts [1, 31, 52, 53] into the CC. This adoption of existing bundles has two motivations: a) to retain intellectual effort – as bundles are often the result of long and costly negotiation (implicit and explicit); and b) to facilitate understanding and automated interaction – as communities and their automated methods will recognise familiar patterns and artifacts.

Nonetheless, the CC cannot be just the union of pre-existing bundles – *harmonisation* (2) plays an essential role. Without harmonisation the CC would be a collection of information silos that preserve domain specific structures together with their boundaries. This would result in a data warehouse that collects data unchanged, thus failing our principal goal that is to facilitate boundary crossing by providing holistic semantic integration.

We harness real *use cases* (3) to tease out and clarify the objectives and aims of the designated communities whose work and communication will be mediated via the CC when they adventure across previous boundaries. To turn an unbounded conceptual space into a manageable space we follow communities' priorities. As use cases evolve and change the associated dependencies and boundaries follow accordingly, thereby identifying required extensions and modifications to the core. Hence, the CC has a clear requirement for flexibility and support for evolution. These guiding principles shape the construction and evolution of the Core Concepts.

4.3. Principles underlying the representation of the Canonical Core

Representation entails metadata, it reflects aspects of the real world for intended purposes and viewpoints [54, 55]. The representation of the CC requires appropriate metadata to describe the complexity of IPC for their supported use cases. As the CC needs to accommodate heterogeneous bundles typically with different encodings, the representation of the core must support what Nilsson called *horizontal harmonisation* [24], that is interoperability across different standards. We adopt the principles for enabling interoperability defined by Duval *et al.* [56], recalled in the Memorandum of Understanding between the Dublin Core Metadata Initiative (DCMI) and the IEEE Learning Technology Standards Committee (IEEE LTSC)[25] and extended by Nilsson *et al.* [57, 24]. These deliver the following:

1. *Extensibility*, ability to create and add new structures to a metadata standard for "application-specific or community-specific needs".

[25] `dublincore.org/documents/2000/12/06/dcmi-ieee-mou`

2. *Modularity*, “ability to combine metadata fragments adhering to different standards”.

3. *Refinements*, “ability to create semantic extensions”.

4. *Multilingualism*, “ability to express, process and display metadata in a number of linguistic and cultural circumstances”.

5. *Machine-processability*, “ability to automate processing of different aspects of the metadata specifications”.

These principles fit the characteristics of an IPC as they assume and acknowledge the co-existence of multiple standards and different specifications. Also, they enable the collaborative approach, for instance members of the designated communities can annotate existing content creating new relationships (1) and refinements (3). Moreover we identify additional issues to consider:

6. *Maturity and level of standardisation* provide a measure of the acceptance among communities as well as an indication of the investments made for uptake. In particular they are reflected in (a) the number of bundles already encoded in a specific representation; (b) the set of available tools compatible with such a representation; and (c) the support offered by communities of experts.

7. *Expressivity and richness* as the ability and the easiness to express logical relationships are important factors that influence the choice of the representation for specific use cases.

8. *Effectiveness* representing the required concepts for the selected application scenario. For instance, verbosity might be more effective in machine-to-machine exchanges whereas terseness might help human reading and understanding (*e.g.* Turtle/RDF[26], N3[27]).

9. *Performance* of the encoding/decoding processes, required to marshall and unmarshall the content of the core. This is an important non-functional engineering aspect that influences the overall behaviour of the system and in particular of the population described in the next section.

10. Support for *validation and consistency* checks. This can be achieved adopting formal restrictions, constraints, description logic, formal rules and inference mechanisms *e.g.* XML Schema, OWL, SHACL, SPARQL-based validation.

[26] `www.w3.org/TR/turtle/`
[27] `www.w3.org/TeamSubmission/n3/`

4.4. Principles underlining the population of the Canonical Core

The population describes the distribution in time of the entities (instances of concepts and instances of relationships between them) in the CC. Population is a dynamic process that is guided by the principles listed below.

1. The *strategy* adopted to populate the CC is influenced by several factors e.g.: volume of data, restrictions, governance, *etc*. However, the possible approaches are: (a) reference or *brokering* – pointers to externally managed bundles are stored in the CC; (b) copy or *harvesting* – the CC holds a physical copy of bundles; and (c) mixed – a combination of the previous two where the CC holds a physical copy of a subset of a bundle, *e.g.* of the information used first or most frequently.

2. Related to the population strategy are the concepts of *conceptual* or *logical* population and *actual* population. The logical population indicates the number of entities which are potentially made available by the CC, whereas the actual population indicates the number of entities currently available. This observation introduces the concept of *latency*, which is the time required to move from the logical to the actual population. For instance, the CC might contain pointers or references to entities of an external catalogue. Although these external entities logically belong to the CC, and thus they are available for the users of the IPC core, there might be a delay to provide access to the concrete objects represented by those entities.

3. *Quality control* is fundamental to manage the population of the core. Quality indicators must be used to assess new entities and providers of entities as well as to modify the population, for instance by removing entities that do not conform to defined quality standards. Pruning, clean-up, deduplication and notification mechanisms can be implemented exploiting such quality indicators.

4. *Governance*, for instance, existing community agreements associated to specific bundles might influence the population strategy and require access control mechanisms.

4.5. Considerations about the boundary regions

In the previous sections we focused our analysis on the characteristics of the CC, we briefly mentioned boundary regions (BR). The CC is an abstraction layer avoiding the complexity of the BR – the core falls under a federation-wide governance whereas BR are independently controlled. For this reason it is difficult to provide a full characterisation of BR. Therefore our focus is at the *interface* between the boundary regions and the core and on the "rules of engagement". Such rules can be modelled leveraging the 'boundary

objects' concept introduced by Star and Griesemer [1, 31]. The authors propose a mechanism to represent and exchange knowledge across organisational borders and facilitate communication. Further CSCW literature built on that concept. Cabitza *et al.* introduce the concept of 'knowledge artifact' that provides *bounded openness*. It "allows participants to establish a shared meaning on the one hand, while remaining open for modifications on the other" [3, 53].

Below we list characteristics of BR that provide the requirements for the interface with the core.

1. BR generate both requirements and constraints for the CC. Such requirements and constraints are time dependent and have a high variation due to the inherent dynamic nature of the regions. Hence, the interface with the core ought to accommodate such *variations*.

2. BR expose a bounded-openness – new boundary regions can be added, removed and at the same time each region can contribute new bundles to the core, provided they fulfil the agreements negotiated with the core.

3. Popular bundles are easily recognised, connected and imported into the core, as they typically gather consensus and form standards whereas less popular bundles constitute extensions. The value of both must be preserved and accounted for, thus the interface has to support both cases and allow differences. In 1945 Vannevar Bush describing memex, wrote "trails that are not frequently followed are prone to fade, items are not fully permanent, memory is transitory" [58]. This captures very well the requirement for promoting and highlighting *extensions* based on diverse criteria in order to engage and attract users and avoid unproductive migrations to other systems, dispersions and so-called "skunk work", where researchers hide their activities to achieve agility and flexibility with consequent loss of evidence for reproducibility and sharing.

To address these requirements the interface between the core and the boundary regions can be modelled as an API for managing extensions. Such an API supports the following operations: 1. registering an extension and holding information about creator and responsible party; 2. noting the aspects of the CC on which the extension depends; 3. winding up work on the extension; and 4. adopting (parts of) the extension into the core.

The following example illustrates how such an API would work in practice. A subcommunity ($SubCom$) harnesses a subset of the CC (C_{sub}) to conduct experimental investigations that yield new data and related concepts, a new conceptual bundle (C_{new}). C_{new}

gains respect and interest from other research groups who would like to use it as early adopters. In order to make it accessible, the API registers C_{new} collecting information about $SubCom$ and C_{sub}. When $SubCom$ has completed the experiments a new (stable) version of C_{new} is available, $C_{new.stable}$. Depending on the relevance or other criteria $C_{new.stable}$ (or parts of it) might be promoted as new bundle in the core. This scenario has implications on the core and calls for additional requirements such as: *versioning* and *provenance*.

5. Building the EPOS Canonical Core

In this section we describe an application of the approach introduced in section 4. We apply our methodology to establish the EPOS CC addressing its three dimensions: definition, representation and population.

5.1. European Plate Observing System (EPOS)

The European Plate Observing System (EPOS)[28] is building a pan-European research infrastructure for solid-Earth sciences. It will start its operational phase in October 2019 with the establishment of an European Research Infrastructure Consortium (ERIC). The mission of EPOS is to integrate the diverse and advanced European Research Infrastructures for solid-Earth sciences creating new opportunities to monitor and understand the dynamic and complex solid-Earth system [59].

EPOS is a prominent example of an IPC that targets ten different scientific communities: seismology, near-fault observatories, GNSS, volcanology, geomagnetic observations, geology, satellite observations, anthropogenic hazards, multi-scale laboratories and geo-energy test beds. It currently involves 141 institutes and organisations spanning 22 countries and connects with the global Earth observation communities. For complexity and scale EPOS provides a rich set of requirements and challenges typical of IPC. Figure 3 depicts a conceptual view of the socio-technical architecture supporting EPOS. Such an architecture is composed by three fundamental elements (from left to right):

- *Thematic Core Services* (TCS) – these are provided by the participating communities. Within each of the targeted domains EPOS has promoted and stimulated the harmonisation of data management, access methods and policies, as well as services (*e.g.* processing, visualisation) and resource provisioning by: 1. fostering the creation of new European-wide thematic hubs; and 2. supporting existing organisations

[28] www.epos-ip.org

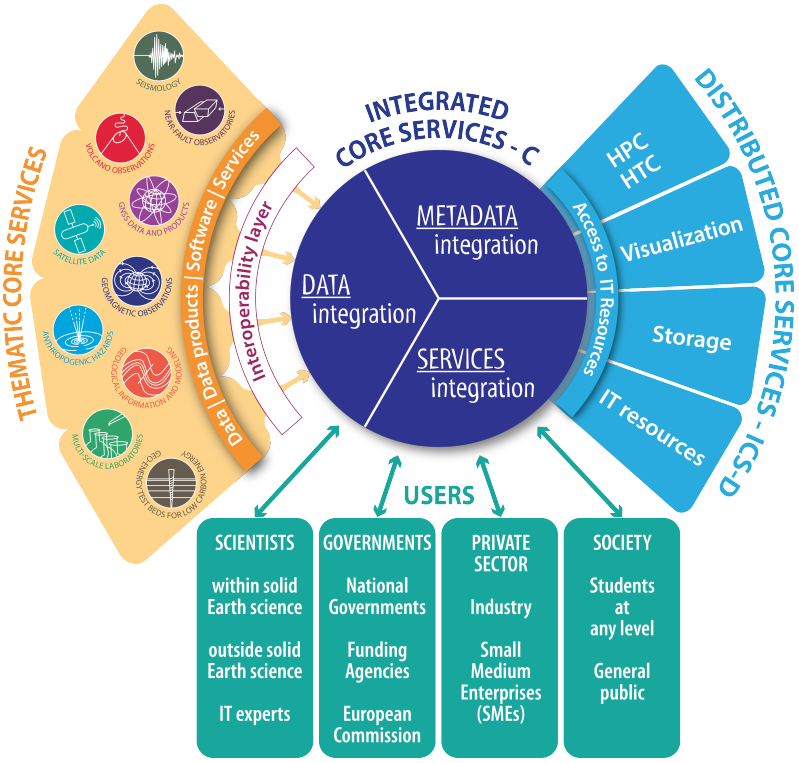


Figure 3: European Plate Observing System architecture high level overview – the EPOS-CC is hosted in the ICS-C. It is represented and maintained in a central metadata catalogue that supports the offered services by steering interactions and information exchanges.

(*e.g.* ORFEUS[29] for seismology). However, much intrinsic diversity remains.

- *Integrated Core Services* - Centralised (ICS-C) – they constitute the novel system under construction to integrate the diverse resources provided by the TCS. Interoperation between the ICS and TCS is needed. This requires the description of available resources by means of rich, flexible and standardised metadata. It supports the data life-cycle from acquisition to exploitation and the conduct of scientific methods and sustained research campaigns.

- *Integrated Core Services* - Distributed (ICS-D) – they constitute the distributed part of the ICS. These services are offered by e-Infrastructure providers and resource providers that – under clear procurement policies or SLAs – make resources available (*e.g.* HPC, HTC, data storage and data transport) for the operation of the ICS's

[29] www.orfeus-eu.org

computational or visualisation tasks.

ICS-C and ICS-D are grouped logically into one component which we refer to as ICS. The metadata describing data and assets are hosted in the EPOS ICS Metadata Catalogue (EIMC). The EPOS CC is represented in the EIMC that underpins the organisation of integration processes and fosters interoperability between the multidisciplinary data, products, software, services and resources of the contributing research communities.

5.2. Definition of the EPOS Canonical Core

The definition of the EPOS CC is conducted by the EPOS metadata group (that includes diverse expertise) based on a set of requirements and use cases collected during the FP7 EPOS-PP (Preparatory Phase) and H2020 EPOS-IP[30] projects, according to the principles presented in 4.2. As EPOS is building an infrastructure on top of existing assets, reuse and adaptation is essential. A specific task was the production of a survey of existing resources contributed by the EPOS designated communities. That survey leveraged: 1. the RIDE database[31]; and 2. reports from focused campaigns with the EPOS communities.

The survey collected information such as providers, contact points, descriptions of resources, and delivered a preliminary classification of the resources in four categories, namely: Data, Data Products, Services and Software (DDSS). Each community contributed a prioritised list of resources to be included in the core based on their maturity and relevance. For instance, the seismological community provided a set of standardised web services[32] (*e.g.* FDSNWS and EIDAWS), primary data (*e.g.* seismic waveform and strong motion data) and data products (*e.g.* earthquake catalogs and hazard maps). Examples of resources by other communities include: InSAR displacement maps, geochemical data, geological maps, meteorological parameters. Figure 4 shows a summary of the current DDSS elements.

The DDSS survey is a valuable asset given the wide scope and heterogeneity of EPOS. To agree it required a strong engagement strategy with the communities exploiting several communication channels. Starting from the DDSS a finer-grained classification has been produced with incremental refinements leading to the definition of the EPOS CC. Such refinements were influenced by geospatial standards (*e.g.* ISO19115) and the CERIF data model [60, 61]. The current EPOS CC includes concepts such as: Dataset, Equipment, Facility, Organisation, Person, Publication, Service, Software, WebService.

[30] https://epos-ip.org/

[31] www.epos-ip.org/ride

[32] http://www.orfeus-eu.org/data/eida/webservices/

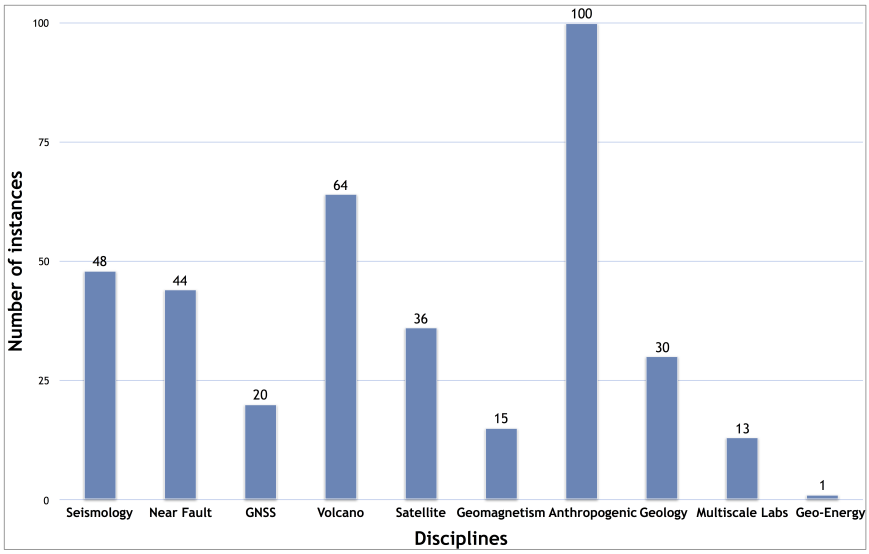
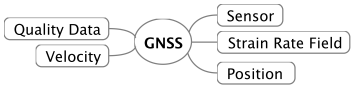


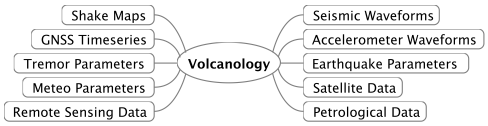
Figure 4: Number of categories of Data, DataProducts, Services and Software (DDSS) offered for sharing by each thematic area

Community bundles (*i.e.* sets of concepts) are made accessible by linking them to the EPOS Core Concepts: *e.g.* $SeismicWaveform \rightarrow Dataset$. Figure 5 offers some examples of such bundles. An overview of the existing resources triggered the *harmonisation* process aimed at providing consistent definitions and interpretations across the EPOS designated communities. Commonalities emerged between diverse disciplines. The DDSS survey revealed overlapping areas across disciplines and highlighted variations in interpretations. For instance, the concept of *Seismic Waveform* is shared across a number disciplines besides Seismology *e.g.* Volcano Observations and Near-Fault Observatories. Similarly the notion of *Event* is quite broadly accepted and in common usage among the communities, however, in some cases there is the need to redefine and/or specialise it – for instance the Anthropogenic Hazards community developed and adopted a slightly different and related definition, which they refer to as an *Episode*. Such examples provide an insight into the typical issues arising in multidisciplinary collaborations. The collaborative work initiated in this process have yielded important results. It has stimulated and encouraged communities to (re)think about their internal knowledge structure, organisation and formalisation. It has fostered the development of shared controlled vocabularies and taxonomies by forming dedicated task forces with a beneficial exchange of expertise. Communities with traditionally more expertise about classifications and knowledge organisation systems, *e.g.* Geology, shared their approaches with communities less experienced in those topics.

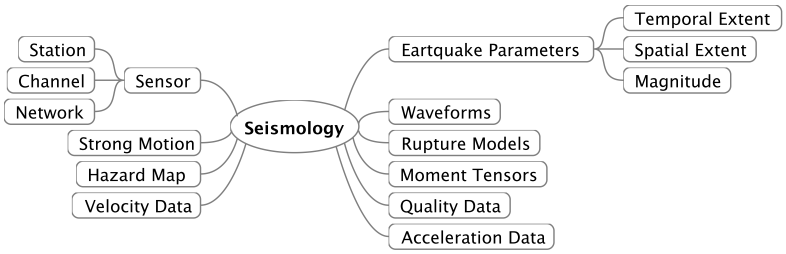
Another important outcome was the identification of representative definitions as well as authoritative sources responsible for each specific bundle and set of entities. This allowed EPOS to avoid duplicated definitions and to provide accurate "reference" definitions with the corresponding representations of the entities. Variations on the reference definitions are allowed where needed but they then need to be linked to and grouped with the reference definitions.



(a) Example of a conceptual bundle from the GNSS community – GNSS concepts can be applied in many contexts *e.g.* to estimate volcano deformations and seismic displacements



(b) Example of a conceptual bundle from the volcano observations community – this community is a predominant example of exploitation of multi-disciplinary, crosscutting concepts



(c) Example of a conceptual bundle from the seismological community

Figure 5: Examples of community bundles – Noteworthy is the presence of overlapping concepts whose definitions might be adopted unaltered by a different community (*e.g.* seismic waveform). However, specialisations, modifications and partial reuse have to be accounted for. In some cases similar concepts may have different interpretations (*e.g.* quality data). The CC has to accommodate diversity and support a range of required scenarios.

The concepts and entities collected in the CC support the use cases and requirements developed by those supporting the IPC and described in section 2.2. The EPOS CC definition is an ongoing process that will continue after EPOS has transitioned to its operational phase [62]. The conceptual framework established and described here will be a valuable tool to support the evolution of this core.

5.3. EPOS Canonical Core representation – EPOS-DCAT-AP

After completing the conceptual definition of the first version of the EPOS CC, the next step was to find a suitable representation that would meet the requirements of the desig-

Encoding	Type	Discipline
MiniSeed	Global	Seismology
WFMetadata-JSON	Community	Seismology
QuakeML	Global	Seismology
Shakemap XML	Community	Seismology
OGC WFS	Global	Geology
OGC WMS	Global	Seismology, Geology
OGC CSW	Global	Geology
CKAN-JSON	Community	Laboratories
Magnetic-HTML	Community	Geo-Magnetic Observatories
OpenSearch XML	Global	Satellite
VpVs-JSON	Community	Near-Fault Observatories
Radon-JSON	Community	Volcanology
CO2-JSON	Community	Volcanology

Table 1: Examples of encodings currently used in EPOS bundles – it provides an overview of the scope and heterogeneity in formats and the adoption of both community and global standards.

nated communities following the principles in 4.3. The CC needed to be formalised in this notation to support a) human communication about the concepts of the core, and b) automated processes assembling, managing, accessing and translating entities corresponding to those concepts.

Along with the overview of the communities' assets, information was collected about the formats, conventions, vocabularies and standards adopted by the communities to represent their resources. In particular the survey revealed that several domain-specific standards co-exists with broader standards. The adoption of standards and shared practices depends on the maturity of the communities. They can be quite heterogeneous. Table 1 provides an example of such diversity. More mature communities follow well-established and broadly applied standards and policies, whereas less mature communities in EPOS need to initiate standardisation and consolidation procedures. The residual inherent heterogeneity is reflected in the composition of the CC and provides additional constraints when choosing a feasible representation. Noteworthy is the adoption of metadata standards for spatial information such as ISO19115, ISO19139, the OGC standards[33] and the INSPIRE conventions [63] *e.g.* by the Geological modelling community.

[33] `http://www.opengeospatial.org/standards`

A representation has been proposed for the EPOS CC building on the DCAT W3C recommendation, namely the EPOS-DCAT-AP. The DCAT data model is represented in RDF, it supports the principles of Linked Open Data (LOD) and reuses concepts from existing vocabularies. Therefore it meets the principles in 4.3. To fulfil the EPOS requirements an EPOS DCAT Application Profile, inspired by Geo-DCAT-AP [20], has been developed extending the general DCAT data model. It follows the recommendation on DCAT-AP extensions [64] and addresses the following concerns:

- Extending the data model with additional concepts required by the EPOS CC (*e.g.* Equipment, Facility, Publication, WebService and Software).
- Introducing new relationships and roles.
- Describing APIs for the programmatic access to datasets.
- Strengthening engagement with scientific communities supporting the inclusion of domain specific knowledge.
- Enabling user-driven approaches and tagging (via annotations).
- Enabling integrity checks and validation (via SHACL).

The latest version of the EPOS-DCAT-AP data model is available online[34] – it includes a UML diagram, ontology definition, examples and more details. AppendixA provides an overview of EPOS-DCAT-AP and its application. Following the DCAT philosophy we reused well-known bundles such as Schema.org and the Web Annotation Vocabulary. When reuse was not possible we created extensions in the EPOS namespace.

The `WebService` entity has been modelled leveraging Schema.org and the Hydra Vocabulary[35] for evolvable Web APIs, a W3C recommendation. This allows us to have flexible and fine grained representations covering the broad EPOS spectrum that includes both global, well-established and community specific standards for web services *e.g.* OGC, FDSN. RDF allows us to include existing domain specific namespaces thus supporting community and user-defined bundles. The `Annotation` entity can be harnessed to enable the collaborative, 'folksonomical' approach – Core Concepts can be enriched with user-driven descriptions and new concepts can be created aggregating, grouping and connecting existing concepts. An important feature is the support for integrity and validation embedded in the representation. This is achieved via the Shapes Constraint Language (SHACL).

[34] `https://github.com/epos-eu/EPOS-DCAT-AP/tree/EPOS-DCAT-AP-shapes`
[35] `http://www.hydra-cg.com/`

It is worth mentioning that the availability of tools that allow representational translation, such as X3ML by FORTH [65], might make the choice of a specific representation less sensitive. Where needed, multiple representations might coexist without affecting the conceptual definitions of the CC.

5.4. Population of the EPOS Canonical Core

Once the EPOS Core Concepts have been identified and agreed, and an appropriate representation chosen, the next step is the population of the CC with real entities from the designated communities. This requires close interaction and collaboration between domain and metadata experts. Ultimately, population needs to be a process that is automated as far as possible. But this requires preparatory work. First experts need to agree the data sources for each concept. They then need to develop import-transformations and protocols. These may stimulate changes at sources and in the CC. Once validated, the parties involved need to agree to sustain the relationships and then an automated process can be coded and run whenever necessary.

To kick off the population process dedicated meetings and workshops were organised targeting the EPOS communities. Documentation, training material, demos and webinars were delivered prior to the face-to-face events in order to inform and prepare the communities for the effort required. This needed to develop the motivation and stimulate the commitment of effort. Moreover, collaborative tools such as wiki and shared repositories[36] have been set up to collect the inputs and feedback from the communities and share documentation and results. To achieve the preliminary population of each community's bundle into the EPOS CC, the communities had to map their resources to the corresponding concepts of the EPOS CC with support from the EPOS-DCAT-AP experts. Due to the scale and complexity of this process the mapping has been carried out in stages prioritising specific entities and adopting in an initial phase a simplified XML representation. Table 2 shows the population of the initial entities.

During the initial population each community uploaded EPOS-DCAT-AP XML compliant files on the EPOS GitHub repository. Those files were successively manually curated and validated. This manual process, albeit costly, helped testing the knowledge collection process and validating the model chosen for the representation. Moreover, it showed an active engagement and participation of the communities who provided useful contributions and feedback. In this phase it has been particularly challenging to keep the alignment of the population with the ongoing refinements of the representation of the core (*i.e.* EPOS-

[36] `https://github.com/epos-eu/EPOS-DCAT-AP`

Entity name	Number of instances
Person	86
Organisation	32
WebService	74

Table 2: Number of instances of the the prioritised entities after the initial (manual) population and validation. In the next stages of the population process (automated) a substantial increase of the number of instances is expected. E.g. Person is expected to grow at least by a factor 100, Organisation by a factor 10 (nearly 260 have been surveyed). The number of instances of Web Service will likely stay in the same order of magnitude and grow at a slower pace. However, in this case it is important to note that the populations made available indirectly by those services are very large.

DCAT-AP). This dynamic situation sometimes introduced issues for the communities, in section 6 we evaluate some those issues with the challenges addressed. According to the principles described in 4.4 one of the goals of the population process is to specify the type of ingestion strategy (1) for each entity (*i.e.* harvesting vs brokering). Therefore the communities have to indicate the requirements and accrual policies associated with their entities – EPOS-DCAT-AP supports this information.

In an operational system the processes of mapping, harvesting and/or brokering of entities are typically delegated to automated methods and tools. To perform the population of the community bundles on a larger scale we devised an architecture with automated components shown in Figure 6. Transformations, convertors and parsers are used to extract the information required by the CC directly from community bundles. SHACL validators[37] help discriminate the admission of entities in the core and debug representation errors. Nevertheless, those technical solutions depend on agreements between the sources of information and the core, commitments to fulfil those agreements and good behaviour – these are essential to ensure that consistent information is delivered.

Examples of mappings of community bundles are available in the EPOS-DCAT-AP GitHub[39].

[37] http://shacl.org

[38] http://epos.cineca.it/apache/mde/public/index.php#

[39] https://github.com/epos-eu/EPOS-DCAT-AP/tree/EPOS-DCAT-AP-shapes/examples

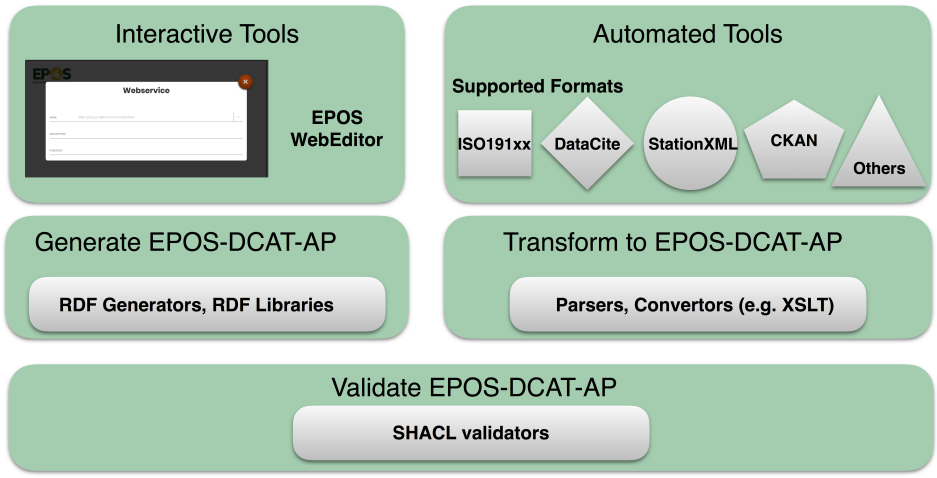


Figure 6: Supporting the population process with automated tools. The diagram indicates the components involved in the population. The ingestion can be performed in an interactive way with the help of a WebEditor[38]. Alternatively a programmatic way is available for a number of supported formats. The converters for each domain specific format are built either reusing existing tools in the case of standards such as ISO191xx, DataCite; or in collaboration with the communities (*e.g.* StationXML). The validation is performed using SHACL validators and the shapes defined for the EPOS model.

6. Evaluation

Providing a complete evaluation of the impact of the presented framework in the EPOS community is unfeasible at this stage due to a number of reasons. The nature and scope of the issues addressed in this research require a longer time scale to be effectively measured. There are individual and organisational aspects that influence adoption and uptake. Those are critical within a single organisation and become much harder in multi-organisational and multi-disciplinary contexts. We target sharing behaviours and working practices that require time to assimilate novel elements. EPOS is currently in its implementation phase [62], for a more complete assessment evaluations ought to be repeated when it is transitioning to its operational phase. These should then be repeated periodically to detect trends. Our framework builds on similar approaches that exploit catalogues and agreed canonical forms in the seismological domain [66]. That experience provided us with useful evidence of benefits and adoption although it has been applied in a more tractable context.

In this section we report an assessment of our work by highlighting some of the challenges encountered and addressed engaging with the EPOS communities. In a recent meet-

ing[40] we asked various key representatives of the EPOS communities, including developers, technical contacts of diverse TCS, domain and metadata experts, leaders and coordinators to provide their feedback about the EPOS-DCAT-AP model. Tables 3 contains the questions and the responses from the participants. Out of approximately 25 participants, who were involved in developing shared knowledge for EPOS, 13 returned responses. In general their responses show a very promising scenario. There are some aspects to improve but it is clear that since the initial presentation of the model to a meeting of four EPOS themes[41] in May 2017, a substantial awareness and understanding have been achieved. The responses to Question 1 show an almost unanimous consensus about the usefulness of the model as a means to facilitate the collection and exchange of domain knowledge.

Table 3: Evaluation survey about EPOS-DCAT-AP

Question	**# of responses**
1: Would the introduction of EPOS-DCAT-AP facilitate the collection and exchange of domain knowledge for the EPOS-ICS?	13
2: Please identify limitations in the proposed EPOS-DCAT-AP. As many as you wish. Where you have suggestions as to how they should be addressed please feel free to make them.	13
3: Are there other contexts that you or your organisation work in where the approach leading to EPOS-DCAT-AP would be useful? Please identify them.	10

Question 2 provides interesting feedback about the perceived limitations of the model. Participants report a number of issues which in some case have been collected in the EPOS-DCAT-AP GitHub. However, most of those issues are related to the initial XML version of the model. In the current RDF version they have been solved. For instance, a better description of `WebService`, building on Schema.org and the Hydra Vocabulary, has been introduced to address previous limitations. Concerns about population strategy and complexity have been addressed by introducing automated tools. A broader set of roles to better support attribution information has been suggested as a possible improvement. We acknowledge the importance of such a requirement and considered to include elements from the PROV vocabulary[42] in order to provide a broader structured provenance information. However, we decided to postpone such a feature to later versions and proposed an intermediate solutions within the current model. One of the answers points out the importance of an agreed strategy for the identification of resources, a feature strongly promoted

[40] www.epos-ip.org/events/epos-implementation-and-validation-workshop-lisbon-portugal12-14-march-2018
[41] https://www.epos-ip.org/milano-and-rome-epos-harmonization-meetings
[42] www.w3.org/TR/prov-o

by EPOS-DCAT-AP – "*The biggest problem (outside of the model) is assigning identifiers to entities and make sure these are consistent [...] We need to agree on rules to set these identifiers and collect them in a (single) repository*". This is an example of an engineering aspect that depends on shared agreements. It shows an increased awareness achieved in the communities about important issues and how our approach stimulated the thinking towards a common shared solution.

Finally, the responses to Question 3 offer the following reflections. A couple of answers identify interesting application contexts. One is positive but shows caution – "*[...] If the EPOS extensions are accepted back into future DCAT standards, this may make implementing EPOS DCAT more attractive*". The remaining answers are more reluctant. For sure at this stage there is still not enough knowledge and trust that would warrant migration from established practices. This is reasonable and in line with the expectations. As already mentioned the introduction of novel elements requires time. Also, local contexts quite often develop solutions tailored to specific needs, the complexity associated with generalisations required in broader contexts can be perceived as an unnecessary overhead. In any case it would be useful to repeat this evaluation when more experience has been acquired and to assess the benefits delivered .

To conclude this analysis, we highlight some key outcomes: the collaborative interaction has been very successful and productive, it allowed us to collect feedback and improve many aspects in order to better support communities' requirements. It encouraged us to think about issues previously unanticipated and developed a common vocabulary and understanding about concepts. This suggests we have a foundation and *modus operandi* for sustainable incremental progress.

7. Conclusions and future work

In this paper we have introduced the concept of Information-Powered Collaborations (IPC), an abstraction the captures the complex dynamics of a modern research context that depends on multi-organisational, multi-disciplinary, multi-national collaboration with increasing complexity and scale. We proposed the formation of an explicit Canonical Core (CC) as their foundation for information sharing and a framework that partitions the complex task of agreeing and maintaining a consistent set of shared Core Concepts to sustain interdisciplinary collaboration. That set has three independent aspects: conceptual definition, representation and population. We have demonstrated how such a framework facilitates the construction and evolution of the information space underpinning an IPC by enabling successive refinements of the three aspects. For instance, communities who are mainly interested in having their entities (*e.g.* data, services and methods) available in the CC will focus

on the population. Those developing automated methods might find the current representation is missing aspects needed and therefore require additions to the representation of the CC. Similarly, someone interested in extending high-level goals might enrich the set of Core Concepts. Thanks to our framework those issues can be addressed independently and progressively, thereby exploiting a separation of concerns. Another important advantage of our framework is that it supports innovation, experiments and heterogeneity. It enables the retention of valued working practices in the Boundary Regions until it is beneficial to transition them into the core, thereby minimising disruption, avoiding constraints and pursuing continuous incremental adoption. Furthermore, it fosters more efficient communication and progressively negotiated agreements between the stakeholders by partitioning the dialogue. As communication is particularly challenging in multi-disciplinary, multi-cultural environments the presented framework provides a significant advance that has been tested in EPOS. We will continue with this approach in EPOS. In particular we plan to:

1. maintain the current set of Core Concepts evolving the Canonical Core when required by new requirements and use cases;

2. further develop and refine the EPOS-DCAT-AP representation, by strengthening the collaboration with the W3C by working with the DXWG in order to make it available for other communities;

3. provide tools leveraging existing components to better support the designated communities in the automated population of their entities. For instance, by means of: graphical interfaces, convertors, mapping services, *etc*; and

4. work on the integration of annotation management tools such as EUDAT B2Note[43] to further exploit the collaborative approach.

Establishing collaborative knowledge to achieve holistic integration and semantic interoperability is an extremely complex task of wide interest that requires alignment of technical, organisational and cultural factors. In order to succeed in this endeavour implications and issues ought to be recognised and addressed effectively, stakeholders acknowledged and good behaviour properly rewarded, *e.g.* by promoting evidence of enhanced scientific results and increasing return on investments. Accommodating local diversity while encouraging migration towards and engagement with the core is essential for sustaining effective collaboration. Although a long way still remains along this path, we believe that the set of principles, the philosophy and the approach proposed are important initial steps.

[43] `https://b2note.bsc.es/`

Acknowledgements

This research has been supported by the following EU projects: EPOS-IP (No. 676564), ENVRI^{plus} (No. 654182), EUDAT2020 (No. 654065), VRE4EIC (No.676247) and DARE (No. 777413). We thank the EPOS WP6-WP7 Team, Andrea Perego from JRC and Daniel Garrjio from ISI USC who provided feedback on the EPOS-DCAT-AP. Finally, we would like to thank two anonymous reviewers for their insightful and constructive comments and suggestions which helped us to improve the manuscript.

Bibliography

[1] S. L. Star, J. R. Griesemer, Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39, Social Studies of Science 19 (3) (1989) 387–420, doi:\bibinfo{doi}{10.1177/030631289019003001}, URL https://doi.org/10.1177/030631289019003001.

[2] M. S. Ackerman, V. Wulf, V. Pipek, Sharing Expertise: Beyond Knowledge Management, MIT Press, Cambridge, MA, USA, ISBN 0262011956, 2002.

[3] M. S. Ackerman, J. Dachtera, V. Pipek, V. Wulf, Sharing knowledge and expertise: The CSCW view of knowledge management, Computer Supported Cooperative Work: CSCW: An International Journal 22 (4-6) (2013) 531–573, ISSN 09259724, doi:\bibinfo{doi}{10.1007/s10606-013-9192-8}.

[4] H. Lubich (Ed.), Towards a CSCW Framework for Scientific Cooperation in Europe, Springer Berlin Heidelberg, ISBN 978-3-540-49115-6, doi:\bibinfo{doi}{https://doi.org/10.1007/3-540-58844-2}, 1995.

[5] B. Fecher, S. Friesike, M. Hebing, What drives academic data sharing?, PLoS ONE 10 (2) (2015) 1–25, ISSN 19326203, doi:\bibinfo{doi}{10.1371/journal.pone.0118053}.

[6] CCSDS. Reference Model for an Open Archival Information System (OAIS). Magenta Book CCSDS 650.0-M-2, 2012. Also published as ISO 14721:2003. Consultative Committee for Space Data Systems., CCSDS - Consultative Committee for Space Data Systems, URL http://public.ccsds.org/publications/archive/650x0m2.pdf., 2012.

[7] H. P. Lubich (Ed.), Foundations of scientific cooperation, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-540-49115-6, 60–73, doi:\bibinfo{doi}{10.1007/

3-540-58844-2_11}, URL https://doi.org/10.1007/3-540-58844-2_11, 1995.

[8] E. Yang, B. Matthews, M. Wilson, Enhancing the core scientific metadata model to incorporate derived data, Future Generation Computer Systems 29 (2) (2013) 612 – 623, ISSN 0167-739X, doi:\bibinfo{doi}{https://doi.org/10.1016/j.future.2011.08.003}, URL http://www.sciencedirect.com/science/article/pii/S0167739X11001427, special section: Recent advances in e-Science.

[9] R. da Silva Machado, R. B. Almeida, D. Y. L. da Rosa, J. L. B. Lopes, A. M. Pernas, A. C. Yamin, EXEHDA-HM: A compositional approach to explore contextual information on hybrid models, Future Generation Computer Systems 73 (2017) 1 – 12, ISSN 0167-739X, doi:\bibinfo{doi}{https://doi.org/10.1016/j.future.2017.03.005}, URL http://www.sciencedirect.com/science/article/pii/S0167739X1630509X.

[10] M. Atkinson, S. Gesing, J. Montagnat, I. Taylor, Scientific workflows: Past, present and future, Future Generation Computer Systems 75 (Supplement C) (2017) 216 – 227, ISSN 0167-739X, doi:\bibinfo{doi}{https://doi.org/10.1016/j.future.2017.05.041}, URL http://www.sciencedirect.com/science/article/pii/S0167739X17311202.

[11] S. J. D. Cox, S. M. Richard, A geologic timescale ontology and service, Earth Science Informatics 8 (1) (2015) 5–19, ISSN 18650481, doi:\bibinfo{doi}{10.1007/s12145-014-0170-6}.

[12] S. J. D. Cox, Time ontology extended for non-Gregorian calendar applications, Semantic Web 7 (2) (2016) 201–209, ISSN 22104968, doi:\bibinfo{doi}{10.3233/SW-150187}.

[13] P. P. Members, Making sense of palaeoclimate sensitivity, Nature 491 (2012) 683, URL http://dx.doi.org/10.1038/nature11574, perspective.

[14] M. Marshal, Ontology spectrum for geological data interoperability, University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC), ISBN 978-90-6164-323-4, 2011.

[15] R. V. Guha, D. Brickley, S. MacBeth, Schema.Org: Evolution of Structured Data on the Web, Queue 13 (9) (2015) 10:10–10:37, ISSN 1542-7730, doi:\bibinfo{doi}{10.

1145/2857274.2857276}, URL http://doi.acm.org/10.1145/2857274.2857276.

[16] European Commission, European data portal, URL www.europeandataportal.eu, last visited on 2017-03-21, 2017.

[17] Open Knowledge International, Swedish open data portal, URL www.opengov.se, last visited on 2017-03-21, 2017.

[18] Open Knowledge Foundation, CKAN, URL https://ckan.org/, last visited on 2017-03-21, 2013.

[19] European Commission, DCAT Application Profile for data portals in Europe Document Metadata, Tech. Rep., European Commission, 2015.

[20] European Commission, GeoDCAT-AP : A geospatial extension for the DCAT application profile for data portals in Europe, Tech. Rep., European Commission, URL https://joinup.ec.europa.eu/node/148281, 2015.

[21] European Commission, StatDCAT-AP – DCAT Application Profile for description of statistical datasets, Tech. Rep., European Commission, 2016.

[22] K. H. Veltman, Syntactic and semantic interoperability: New approaches to knowledge and the semantic web, New Review of Information Networking 7 (1) (2001) 159–183, ISSN 1361-4576, doi:\bibinfo{doi}{10.1080/13614570109516975}.

[23] L. M. Chan, M. L. Zeng, Metadata interoperability and standardization – A study of methodology part I: Achieving interoperability at the schema level, D-Lib Magazine 12 (6) (2006) 23–41, ISSN 10829873, doi:\bibinfo{doi}{10.1111/j.1468-0084.2006.00151.x}.

[24] M. Nilsson, From Interoperability to Harmonization in Metadata Standardization: Designing an Evolvable Framework for Metadata Harmonization, Ph.D. thesis, KTH, Stockholm, URL http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-26057, 2010.

[25] G. Alemu, B. Stevens, P. Ross, Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries, New Library World 113 (1/2) (2012) 38–54, ISSN 0307-4803, doi:\bibinfo{doi}{10.1108/03074801211199031}.

[26] B. Haslhofer, W. Klas, A survey of techniques for achieving metadata interoperability, ACM Computing Surveys 42 (2) (2010) 1–37, ISSN 03600300, doi:

\bibinfo{doi}{10.1145/1667062.1667064}, URL http://portal.acm.org/citation.cfm?doid=1667062.1667064.

[27] Shirky, Clay, Ontology is overrated: categories, links, and tags, URL www.shirky.com/writings/ontology_overrated.html, last visited on 2017-06-02, 2005.

[28] V. Wal, Thomas, Folksonomy: coinage and definition, URL http://vanderwal.net/folksonomy.html, last visited on 2017-06-02, 2007.

[29] T. Gruber, Ontology of Folksonomy: A Mash-Up of Apples and Oranges, Int. J. Semantic Web Inf. Syst. 3 (2007) 1–11.

[30] G. Antoniou, F. van Harmelen, Web Ontology Language: OWL, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-540-24750-0, 67–92, doi:\bibinfo{doi}{10.1007/978-3-540-24750-0_4}, URL https://doi.org/10.1007/978-3-540-24750-0_4, 2004.

[31] S. L. Star, This is Not a Boundary Object: Reflections on the Origin of a Concept, Science, Technology, & Human Values 35 (5) (2010) 601–617, doi:\bibinfo{doi}{10.1177/0162243910377624}, URL https://doi.org/10.1177/0162243910377624.

[32] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. a.C 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. a. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018, ISSN 2052-4463, doi:\bibinfo{doi}{10.1038/sdata.2016.18}, URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4792175{\&}tool=pmcentrez{\&}rendertype=abstract.

[33] L. Candela, D. Castelli, P. Pagano, Virtual Research Environments : An Overview and a Research Agenda 12 (August) (2013) 75–81.

[34] M. Assante, L. Candela, D. Castelli, G. Coro, L. Lelii, P. Pagano, Virtual research environments as-a-service by gCube, CEUR Workshop Proceedings 1871 (June) (2016) 8–10, ISSN 16130073, doi:\bibinfo{doi}{10.7287/peerj.preprints.2511v1}.

[35] S. Gesing, N. Wilkins-Diehr, Science gateway workshops 2014 special issue conference publications, Concurrency and Computation: Practice and Experience 27 (16) (2015) 4247–4251, ISSN 1532-0634, doi:\bibinfo{doi}{10.1002/cpe.3615}, URL http://dx.doi.org/10.1002/cpe.3615.

[36] M. Agosti, N. Ferro, G. Silvello, Digital library interoperability at high level of abstraction, Future Generation Computer Systems 55 (2016) 129 – 146, ISSN 0167-739X, doi:\bibinfo{doi}{http://dx.doi.org/10.1016/j.future.2015.09.020}, URL http://www.sciencedirect.com/science/article/pii/S0167739X15003003.

[37] National Research Council, Astronomy and Astrophysics in the New Millennium, The National Academies Press, Washington, DC, ISBN 978-0-309-07031-7, doi:\bibinfo{doi}{10.17226/9839}, URL https://www.nap.edu/catalog/9839/astronomy-and-astrophysics-in-the-new-millennium, 2001.

[38] R. Hanisch, G. Berriman, T. Lazio, S. E. Bunn, J. Evans, T. McGlynn, R. Plante, The Virtual Astronomical Observatory: Re-engineering access to astronomical data, Astronomy and Computing 11 (2015) 190 – 209, ISSN 2213-1337, doi:\bibinfo{doi}{https://doi.org/10.1016/j.ascom.2015.03.007}, URL http://www.sciencedirect.com/science/article/pii/S2213133715000256, the Virtual Observatory: II.

[39] R. Hanisch, The Virtual Observatory: I, Astronomy and Computing 7-8 (2014) 1 – 2, ISSN 2213-1337, doi:\bibinfo{doi}{https://doi.org/10.1016/j.ascom.2014.07.006}, URL http://www.sciencedirect.com/science/article/pii/S2213133714000341, special Issue on The Virtual Observatory: I.

[40] D. van Uytvanck, H. Stehouwer, L. Lampen, Semantic metadata mapping in practice: the Virtual Language Observatory, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) (2012) 1029–1034.

[41] T. Tiropanis, W. Hall, N. Shadbolt, D. De Roure, N. Contractor, J. Hendler, The web science observatory, IEEE Intelligent Systems 28 (2) (2013) 100–104, ISSN 15411672, doi:\bibinfo{doi}{10.1109/MIS.2013.50}.

[42] T. Tiropanis, W. Hall, J. Hendler, C. de Larrinaga, The Web Observatory: A Middle Layer for Broad Data, Big Data 2 (3) (2014) 129–133, doi:\bibinfo{doi}{10.1089/big.2014.0035}.

[43] T. Tiropanis, X. Wang, R. Tinati, W. Hall, Building a Connected Web Observatory Architecture and Challenges, 2nd International Workshop on Building Web Observatories (B-WOW14), ACM Web Science Conference 2014 .

[44] X. Wang, T. Tiropanis, R. Tinati, WDFed: Exploiting Cloud Databases Using Metadata and RESTful APIs, Springer International Publishing, ISBN 978-3-319-49157-8, 345–356, doi:\bibinfo{doi}{10.1007/978-3-319-49157-8_30}, 2016.

[45] M. Jamshidi, System of Systems Engineering: Innovations for the 21st Century, ISBN 9780470195901, doi:\bibinfo{doi}{10.1002/9780470403501}, 2008.

[46] S. Nativi, P. Mazzetti, M. Santoro, F. Papeschi, M. Craglia, O. Ochiai, Big Data challenges in building the Global Earth Observation System of Systems, Environmental Modelling and Software 68 (2015) 1–26, ISSN 13648152, doi:\bibinfo{doi}{10.1016/j.envsoft.2015.01.017}, URL http://dx.doi.org/10.1016/j.envsoft.2015.01.017.

[47] M. Santoro, S. Nativi, P. Mazzetti, Contributing to the GEO Model Web implementation: A brokering service for business processes, Environmental Modelling and Software 84 (2016) 18–34, ISSN 13648152, doi:\bibinfo{doi}{10.1016/j.envsoft.2016.06.010}, URL http://dx.doi.org/10.1016/j.envsoft.2016.06.010.

[48] N. Chungoora, R. I. Young, G. Gunendran, C. Palmer, Z. Usman, N. A. Anjum, A.-F. Cutting-Decelle, J. A. Harding, K. Case, A model-driven ontology approach for manufacturing system interoperability and knowledge sharing, Computers in Industry 64 (4) (2013) 392 – 401, ISSN 0166-3615, doi:\bibinfo{doi}{https://doi.org/10.1016/j.compind.2013.01.003}, URL http://www.sciencedirect.com/science/article/pii/S0166361513000055.

[49] A. L. Szejka, O. C. Junior, The Application of Reference Ontologies for Semantic Interoperability in an Integrated Product Development Process in Smart Factories, Procedia Manufacturing 11 (2017) 1375 – 1384, ISSN 2351-9789, doi:\bibinfo{doi}{https://doi.org/10.1016/j.promfg.2017.07.267}, URL http://www.sciencedirect.com/science/article/pii/

S2351978917304754, 27th International Conference on Flexible Automation and Intelligent Manufacturing, FAIM2017, 27-30 June 2017, Modena, Italy.

[50] M. Imran, R. Young, Reference ontologies for interoperability across multiple assembly systems, International Journal of Production Research 54 (18) (2016) 5381–5403, doi:\bibinfo{doi}{10.1080/00207543.2015.1087654}, URL https://doi.org/10.1080/00207543.2015.1087654.

[51] European Commission, Open data An engine for innovation, growth and transparent governance Swedish open data portal – COM(2011) 882 final, URL http://www.europarl.europa.eu/RegData/docs_autres_institutions/commission_europeenne/com/2011/0882/COM_COM(2011)0882_EN.pdf, last visited on 2017-11-23, 2011.

[52] F. Cabitza, C. Simone, M. Sarini, Knowledge Artifacts as Bridges between Theory and Practice: The Clinical Pathway Case, Springer US, Boston, MA, ISBN 978-0-387-09659-9, 37–50, doi:\bibinfo{doi}{10.1007/978-0-387-09659-9_3}, URL http://dx.doi.org/10.1007/978-0-387-09659-9_3, 2008.

[53] F. Cabitza, G. Colombo, C. Simone, Leveraging underspecification in knowledge artifacts to foster collaborative activities in professional communities, International Journal of Human-Computer Studies 71 (1) (2013) 24 – 45, ISSN 1071-5819, doi:\bibinfo{doi}{http://doi.org/10.1016/j.ijhcs.2012.02.005}, URL http://www.sciencedirect.com/science/article/pii/S107158191200033X, special Issue on supporting shared representations in collaborative activities.

[54] G. Alemu, B. Stevens, 1 - Introduction, in: G. Alemu, B. Stevens (Eds.), An Emergent Theory of Digital Library Metadata, Chandos Publishing, ISBN 978-0-08-100385-5, 1 – 9, doi:\bibinfo{doi}{https://doi.org/10.1016/B978-0-08-100385-5.00001-8}, URL http://www.sciencedirect.com/science/article/pii/B9780081003855000018, 2015.

[55] R. Gartner, What Metadata Is and Why It Matters, Springer International Publishing, ISBN 978-3-319-40893-4, 1–13, doi:\bibinfo{doi}{10.1007/978-3-319-40893-4_1}, URL http://dx.doi.org/10.1007/978-3-319-40893-4_1, 2016.

[56] E. Duval, W. Hodgins, S. Sutton, S. L. Weibel, Metadata principles and practicalities, D-Lib Magazine 8 (4) (2002) 1–15, ISSN 10829873, doi:\bibinfo{doi}{10.1045/april2002-weibel}.

[57] M. Nilsson, P. Johnston, Towards an interoperability framework for metadata standards, International Conference on Dublin Core and Metadata Applications (March 2005), URL http://dcpapers.dublincore.org/index.php/pubs/article/viewArticle/835.

[58] V. Bush, J. Wang, As we may think, Atlantic Monthly 176 (1945) 101–108.

[59] D. Bailo, R. Paciello, R. Rabissoni, M. Sbarra, V. Vinciarelli, Integration of heterogeneous data, software and services in Solid Earth Sciences: the EPOS system design and roadmap for the building of Integrated Core Services., Tech. Rep., INGV, 2018.

[60] K. G. Jeffery, D. Bailo, EPOS: Using Metadata in Geoscience, Springer International Publishing, Cham, ISBN 978-3-319-13674-5, 170–184, doi:\bibinfo{doi}{10.1007/978-3-319-13674-5_17}, URL https://doi.org/10.1007/978-3-319-13674-5_17, 2014.

[61] D. Bailo, D. Ulbricht, M. L. Nayembil, L. Trani, A. Spinuso, K. G. Jeffery, Mapping Solid Earth Data and Research Infrastructures to CERIF, Procedia Computer Science 106 (2017) 112 – 121, ISSN 1877-0509, doi:\bibinfo{doi}{dx.doi.org/10.1016/j.procs.2017.03.043}, URL http://www.sciencedirect.com/science/article/pii/S1877050917303113.

[62] M. Cocco, EPOS IP Management Plan D1.1, doi:\bibinfo{doi}{10.5281/zenodo.1213698}, URL https://doi.org/10.5281/zenodo.1213698, 2018.

[63] EU Parliament, Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), Official Journal of the European Union 50 (L108).

[64] PwC EU Services, Analysis of the DCAT-AP extensions, Tech. Rep., European Commission, URL https://joinup.ec.europa.eu/sites/default/files/document/2017-10/DCAT-APextensionsanalysis_v1.00.pdf, 2017.

[65] N. Minadakis, Y. Marketakis, H. Kondylakis, G. Flouris, M. Theodoridou, M. Doerr, G. de Jong, X3ML Framework: An effective suite for supporting data mappings (2015) 1–12 19th International Conference on Theory and Practice of Digital Libraries (TPDL2015).

[66] L. Trani, M. Koymans, M. Atkinson, R. Sleeman, R. Filgueira, WFCatalog: A catalogue for seismological waveform data, Computers & Geosciences 106

(2017) 101 – 108, ISSN 0098-3004, doi:\bibinfo{doi}{https://doi.org/10.1016/j.cageo.2017.06.008}, URL http://www.sciencedirect.com/science/article/pii/S0098300416308263.

AppendixA. EPOS-DCAT-AP and examples of its application

In this appendix we present a simplified UML class diagram of the EPOS-DCAT-AP model and examples of encodings in the RDF/Turtle notation. Figure A.7 shows details of the extensions built on top of DCAT-AP v1.1[44]. The additional classes introduced are represented in yellow. They allow us to address the specific requirements of the EPOS community. In particular, they enable the description of additional concepts beyond Dataset (the main focus of DCAT). For instance, Service and WebService allow the mapping of important community assets. Such concepts can be included in a catalogue with: $Catalog \xrightarrow{\text{epos:resource}} Resource$ as illustrated in Listing 1.

Web Services are very important in EPOS as they provide programmatic access to a variety of datasets and resources. In EPOS-DCAT-AP we are able to describe such a programmatic access by harnessing the relationships $Distribution \xrightarrow{\text{dct:conformsTo}} WebService$ and $Distribution \xrightarrow{\text{dcat:accessURL}} Operation$.

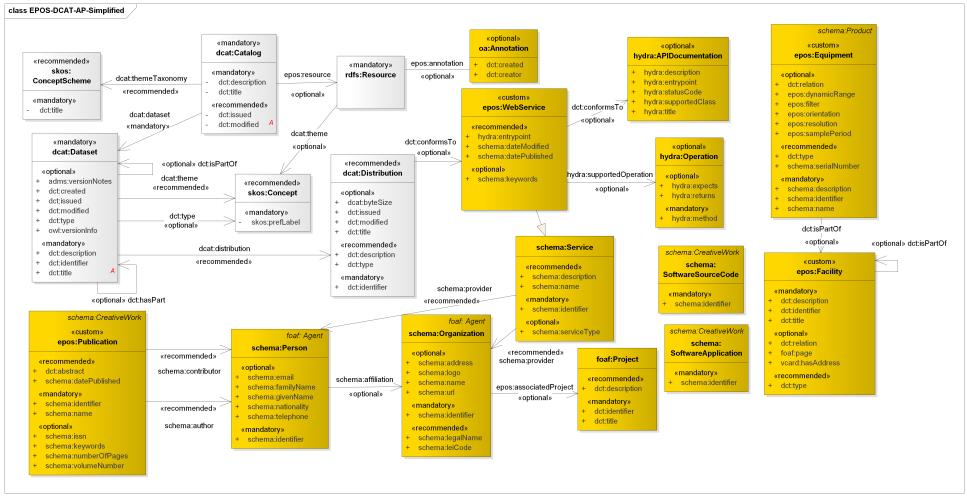


Figure A.7: Simplified UML model of EPOS-DCAT-AP. We extended DCAT-AP v1.1 with additional classes (in yellow) and their relationships in order to fulfill the EPOS requirements

Listing 1 provides an example of catalogue including different resources.

[44] https://joinup.ec.europa.eu/release/dcat-ap-v11

Listing 1: It shows the start of the definition of the conceptual space, `ConceptScheme`, for EPOS, `Epos`. The established namespaces from which terms are imported are defined. The concept `catalogID` is then introduced and the first five attributes of its elements to hold metadata are defined. Others are omitted. Their resources, `resource`, are then defined, but for clarity their details are omitted. A resource in this context is an asset relevant for EPOS, e.g. a Facility, an Equipment and a Web Service. Format is RDF/Turtle.

```
@prefix epos: <http://www.epos−eu.org/epos−dcat−ap#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
## A scheme that includes EPOS concepts
  <epos:Epos> a skos:ConceptScheme;
     dct:title ‘‘EPOS concepts”@en;
 dct:description ‘‘It contains the concepts of the EPOS domain”@en; .
## A catalogue that collects EPOS assets
  <catalogID>  a dcat:Catalog;
 dct:title ‘‘EPOS Metadata Catalogue”@en;
 dct:description ‘‘A catalogue that represents the EPOS CC”@en;
 ## A skos taxonomy of the domain specific concepts in EPOS
 dcat:themeTaxonomy <epos:Epos>;
 ## Including datasets and data collections
 dcat:dataset <datasetID>;
 dcat:dataset <datasetCollectionID>;
 ...
 ## Including additional assets
 epos:resource <equipmentID>;
 epos:resource <facilityID>;
 epos:resource <webserviceID>
 ... .
```

Instruments can be described with the entity `Equipment` and linked to a specific `Facility` $Equipment \xrightarrow{\text{dct:isPartOf}} Facility$. Listing 2 provides an example of a seismic network described as a `epos:Facility` with a seismic station (`epos:Equipment`) and a seismic stream (channel) (`epos:Equipment`), such a representation is obtained by mapping a seismological standard: StationXML[45].

[45] `www.fdsn.org/xml/station/`

Listing 2: Example of mappings of Facility and Equipment. A classification of seismological concepts is defined. It includes the concepts `SeismicNetwork` and `SeismicStation`. Then instances of concepts, `Facility` and `Equipment`, are defined with some of their attributes. Others are omitted for clarity. Format is RDF/Turtle.

```
## A classification of seismological concepts
  <epos:Seismology> a skos:ConceptScheme;
    dct:title ‘‘Seismology”@en;
    dct:description ‘‘It contains the concepts of the Seismology domain”@en; .
## Defining the concept Seismic Network
  <epos:SeismicNetwork> a skos:Concept ;
    skos:definition ‘‘Collection of seismic stations in a seismic network”;
    skos:inScheme <Seismology> ;
    skos:prefLabel ”Seismic Network” .
## Defining the concept Seismic Station
  <epos:SeismicStation> a skos:Concept ;
    skos:definition ‘‘A station for recording oscillations of the Earth’s surface”;
    skos:inScheme <Seismology> ;
    skos:prefLabel ‘‘Seismic Station”
    skos:altLabel ‘‘Seismometer”.
## Describing a seismic network (NL) as a Facility
  <EPOS/ORFEUS/EIDA/ODC/NL> a epos:Facility ;
    dct:description ‘‘Netherlands Seismic and Acoustic Network”;
    ##Seismological networks follow the FDSN recommendation to adopt
    ##DOIs (www.fdsn.org/services/doi/)
    dct:identifier <doi.org/10.21944/e970fd34−23b9−3411−b366−e4f72877d2c5> ;
    dct:title ‘‘Seismic Network NL”;
    dcat:contactPoint <ContactID> ;
    ##The concept associated to this Equipment
    dcat:theme <SeismicNetwork>;
    ... .
## Describing a seismic station as an Equipment
  <EPOS/ORFEUS/EIDA/ODC/NL.HGN> a epos:Equipment ;
    dct:description ‘‘Broadband Seismic Station HEIMANSGROEVE, NETHERLANDS ” ;
    dct:identifier <EPOS/ORFEUS/EIDA/ODC/NL.HGN> ;
    ## Location
    dct:spatial [ a dct:Location ;
      locn:geometry ‘‘POINT(50.764 5.9317 135.0)” ] ;
    dct:title ‘‘Seismic Station NL.HGN”;
    ##The concept associated with this Equipment
    dcat:theme <SeismicStation>;
    ## This station belongs to the NL network
    dct:isPartOf <EPOS/ORFEUS/EIDA/ODC/NL>;
    ... .
```

```
## A seismic stream belonging to a station
  </EPOS/ORFEUS/EIDA/ODC/NL.HGN.02.BHZ> a epos:Equipment ;
    dct:description ‘‘Seismic stream recording ground motion”;
    dct:identifier <EPOS/ORFEUS/EIDA/ODC/NL.HGN.02.BHZ>;
    ## This stream belongs to the NL.HGN station
    dct:isPartOf <EPOS/ORFEUS/EIDA/ODC/NL.HGN> ;
    dct:spatial [ a dct:Location ;
      locn:geometry ”POINT(50.764 5.9317 135.0)”] ;
    dct:title ‘‘Seismic Stream NL.HGN.02.BHZ”;
    epos:orientation ”0.0/−90.0”;
    epos:samplePeriod ”0.025”;
    ... .
```

The Listing 3 shows an example of classification using SKOS. It can be used to describe domain specific concepts (*e.g.* Seismic Waveform) which can be associated with the EPOS Core Concepts *e.g.* Dataset, Webservice: $Resource \xrightarrow{\text{dcat:theme}} Concept$.

Listing 3: Example of classification using SKOS. It groups knowledge in concept schemes, `ConceptScheme`. Here we see a few members, `Concept`, be gathered under the `Seismology` theme's heading, and then a group of concepts being gathered under the `VolcanoObservations` heading, with one concept, `SeismicWaveform` shared. Format is RDF/Turtle.

```
@prefix epos: <http://www.epos−eu.org/epos−dcat−ap#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#>.
@prefix owl: <http://www.w3.org/2002/07/owl#> .
#####
##### Example of a classification of domain specific concepts to be associated with the EPOS CC
#####
## Communities can define and manage their sets of concepts in a concept scheme
<epos:Seismology> a skos:ConceptScheme;
    dct:title ‘‘Seismology”@en;
 dct:description ‘‘It contains the concepts of the Seismology domain”@en; .
## Defining the concept SeismicWaveform with multi−lingual support
<epos:SeismicWaveform> a skos:Concept;
 skos:definition ‘‘Measurement of the dynamic displacement of the Earth”@en;
 skos:inScheme <epos:Seismology>;
 skos:prefLabel ‘‘Seismic waveform”@en;
 skos:prefLabel ‘‘Forma d’onda sismica”@it;
 #can be used by applications for text−based indexing/search (e.g. via a web interface)
 skos:hiddenLabel ‘‘seismic_waveform”@en;
    skos:hiddenLabel ‘‘MSEED”@en; .
## Another seismological concept
<epos:SeismicHazardMap> a skos:Concept;
```

```
 skos:definition ‘‘A map that shows the hazard associated with potential earthquakes in a particular area”;
 skos:inScheme <epos:Seismology>;
 skos:prefLabel ‘‘Seismic hazard map”@en ;
 skos:altLabel ‘‘Seismological hazard map”@en; .
<epos:VolcanoObservations> a skos:ConceptScheme;
     dct:title ‘‘VolcanoObservations”@en;
 dct:description ‘‘It contains the concepts of the Volcano Observations”@en;  .
<epos:GeochemicalData> a skos:Concept;
 skos:definition ‘‘It refers to the types of geochemical  ....”@en;
 skos:inScheme <epos:VolcanoObservations>;
 skos:prefLabel ‘‘Geochemical Data”@en;
     skos:altLabel ‘‘Geochemistry”@en; .
## SeismicWaveform belongs to more that one concept schemes, i.e. it is a shared concept
<epos:SeismicWaveform> a skos:Concept;
 skos:definition ‘‘Measurement of the dynamic displacement of the Earth”@en;
 skos:inScheme <epos:VolcanoObservations>;
 skos:prefLabel ‘‘Seismic waveform”@en;
 skos:altLabel ‘‘Seismology”@en; .
## Importing an existing ontology. Communities who already invested in the definition
## of formalised knowledge can retain their investments.
<CommunityOntology> a owl:Ontology, skos:ConceptScheme .
```

For more details and examples we refer readers to the online documentation[46].

[46] https://github.com/epos-eu/EPOS-DCAT-AP/