"FAIR friendly research data catalogues: How far are we? "

3 April 2017, Barcelona, Spain

1. Introduction

Back in 2014, a diverse set of stakeholders - representing academia, industry, funding agencies, and scholarly publishers - have come together to design and jointly endorse a concise and measureable set of principles known as **FAIR Data Principles**, i.e. principles aiming at facilitating humans and / or machines in the "(re-)use" of data by making such data Findable, Accessible, Interoperable, and Reusable. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data. The aim and focus of the BlueBRIDGE[1] project workshop organised in Barcelona on Monday 3 April 2017 (co-located with the 9th RDA Plenary Meeting[2]) was precisely on looking at: **"FAIR friendly research data catalogues: How far are we?"[3] to really understand how many of the existing research data catalogues are FAIR-friendly.**

The first report from the High-Level Expert Group (HLEG) on the European Open Science Cloud also endorsed this vision. The HLEG recommends "framing the EOSC as the EU contribution to a future, global Internet of FAIR Data and Services underpinned by open protocols"[4] . The EOSC will have to support the Finding, Access, Interoperability and in particular the **Re-use of open, as well as sensitive and properly secured data.**

It will also have to support the data related elements (software, standards, protocols, workflows) that enable re-use and data driven knowledge discovery and innovation.

[1] www.bluebridge-vres.eu

[2] https://www.rd-alliance.org/plenaries/rda-ninth-plenary-meeting-barcelona/rda-9th-plenary-programme

[3] www.bluebridge-vres.eu/events/bluebridge-workshop-fair-friendly-research-data-catalogues-how-far-are-we-3-april-2017

[4] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

The workshop brought together **over 40 representatives** of H2020 projects, e-infrastructures, European and global initiatives and data users currently dealing with research data catalogues (See the list of participants[5]). Discussions focused on how these initiatives are approaching the FAIR principles, what status they are at and how they plan to move forward.

Figure 1: Workshop participants

[5] http://www.bluebridge-vres.eu/FAIR-workshop/participants

All the participants believed that the workshop was very productive as it opened a constructive dialogue between different initiatives and countries. Such dialogue is fundamental to achieve consensus on a common vision on FAIR. This is why, some of the organisers and attendees participating in the EOSCpilot project[6], have therefore decided to submit the proposal for a follow-up of the BlueBRIDGE Barcelona workshop within the Open Science Fair conference[7], 6-8 September 2017 in Athens.

2. Workshop Agenda

13:30 - 14:00	*Welcome Coffee*
14:00 - 14:05	**Welcome & Overview of the workshop objectives**, *Donatella Castelli, CNR-ISTI & BlueBRIDGE Coordinator*
14:05 - 14:25	**A research data catalogue supporting Blue Growth: the BlueBRIDGE case,** *Massimiliano Assante, CNR-ISTI*
14:25 - 15:05	**How existing Research Infrastructures are dealing with data catalogues: the ICOS & EPOS experiences** • **ICOS: Integrated Carbon Observation System**, *Harry Lankreijer, Lund University* • **European Plate Observing System (EPOS) metadata driven catalogue and its synergy with other initiatives (VRE4EIC and others)**, *Daniele Bailo, INGV*
15:05 - 15:30	*Interactive Discussion*
15:30 - 16:00	***Coffee Break***
16:00 - 16:40	**Research data catalogues supporting life & earth sciences: the approaches adopted by ELIXIR & Copernicus** • **Research data catalogues and data interoperability in life sciences**, *Rafael Jimenez, ELIXIR* • **How FAIR is GEOSS**, *Mattia Santoro, CNR-IIA*
16:40 - 17:00	**EUDAT-B2FIND: A FAIR-friendly and interdisciplinary data catalogue,** *Heinrich Widmann, German Climate Computing Center (DKRZ)*
17:00 - 17:20	**An ecosystem to support FAIR Data**, *Luiz Olavo Bonino, Dutch Techcentre & Vrije University Amsterdam*
17:20 - 18:00	*Interactive Discussion*
18:00 - 18:45	***Networking Cocktail***

All the workshop presentations are available on the BlueBRIDGE website[8].

[6] https://eoscpilot.eu/
[7] https://eoscpilot.eu/events/open-science-fair-6-8-september-2017-athens
[8] http://www.bluebridge-vres.eu/FAIR-workshop/agenda

3. How many of the existing research data catalogues are FAIR-friendly?

Seven presentations illustrating how the different data catalogues are approaching the FAIR principles were given in the workshop and they are reported in the following section.

A research data catalogue supporting Blue Growth: the BlueBRIDGE case, Massimiliano Assante, ISTI-CNR[9]

BlueBRIDGE is an H2020 project delivering specialised data services for aquaculture farm management, the ecosystem approach to fisheries, spatial data analysis, and marine spatial planning. These services are operated through Virtual Research Environments[10] built on top of a hybrid-data infrastructure (D4Science)[11]. This hybrid-data infrastructure is capable of dynamically federating computing and storage resources coming from third party providers, datasets belonging to heterogeneous data sources, analytical tools developed by different organisations, and to offer all of these services in a unique collaborative environment.
The BlueBRIDGE data catalogue[12] has been designed with the FAIR principles in mind. The approach adopted in the context of the BlueBRIDGE project, culminates in the design and implementation of an open, flexible and rich Research Data Catalogue where a large set of heterogeneous research products can be (i) published, (ii) seamlessly discovered and (iii) accessed from users and dedicated services. Currently, the catalogue contains resources suitable for and resulting from the services and virtual research environments (VREs) operated by the BlueBRIDGE consortium to serve cases ranging from stock assessment to aquaculture atlas generation, strategic investment and scientific training. Datasets include, among the others, species distribution maps, environmental data, statistical data and area regulation zones. All the products are accompanied with rich descriptions capturing general attributes, e.g. title and creator(s), as well as usage policies and licences. The approach adopted by BlueBRIDGE proposes to exploit the capabilities offered by the D4Science hybrid data infrastructure. D4Science enacts the catalogue FAIRness by deploying and operating a set of services and facilities that enable to access the catalogue items payload (beyond metadata) in the context where these items were produced, thus by overcoming interoperability and reusability issues.

ICOS: Integrated Carbon Observation System, Harry Lankreijer, Lund University[13]

ICOS[14] is a pan-European research infrastructure (RI) for observing and understanding the greenhouse gas (GHG) balance of Europe and its adjacent regions. The major task of ICOS is

[9] https://www.slideshare.net/BlueBridgeVREs/a-research-data-catalogue-supporting-blue-growth-the-bluebridge-case
[10] http://www.bluebridge-vres.eu/news/vre-three-letters-enhance-cooperation-and-collaboration-amongst-researchers-modern-science
[11] www.d4science.org
[12] https://bluebridge.d4science.org/catalogue
[13] www.slideshare.net/BlueBridgeVREs/icos-integrated-carbon-observation-system-open-data-to-open-our-eyes-to-climate-change
[14] www.icos-ri.eu/

to collect and make available in a transparent manner, the high-quality observational data from its state-of-the-art measurement stations. These ICOS data – from atmosphere, ecosystem and ocean stations – will contribute to research aiming to describe and understand the present state of the global carbon cycle. The Carbon Portal will be the virtual data center that present the data products and make it available.
The work of ICOS and the Carbon Portal towards open data with FAIR principles is described in the following: ICOS has an open data policy with free use, requesting the user to give appropriate credit (Creative Commons Attribution 4.0). The Carbon Portal is developing a data catalogue using an ontology based on a semantic metadata description. This will make it possible to integrate ICOS observations with data from other RI's as well with data of global networks. For integration, the Carbon Portal is actively following the developments of international standards for eg. metadata and data citation. The development of easy editable metadata schemes and of good systems for data citation should contribute to motivate researchers to publish their data.

European Plate Observing System (EPOS) metadata driven catalogue and its synergy with other initiatives (VRE4EIC and others), Daniele Bailo, INGV[15]

EPOS[16], the European Plate Observing System, has been designed with the vision of creating a pan-European infrastructure for solid Earth science to support a safe and sustainable society. In accordance with this scientific vision, the EPOS mission is to integrate the diverse and advanced European Research Infrastructures for solid Earth science relying on new e-science opportunities to monitor and unravel the dynamic and complex Earth System. EPOS will enable innovative multidisciplinary research for a better understanding of the Earth's physical and chemical processes that control earthquakes, volcanic eruptions, ground instability and tsunami as well as the processes driving tectonics and Earth's surface dynamics. To accomplish its mission, EPOS is engaging different stakeholders, not limited to scientists, to allow the Earth sciences to open new horizons in our understanding of the planet. EPOS also aims at contributing to prepare society for geo-hazards and to responsibly manage the exploitation of geo-resources. Through integration of data, models and facilities, EPOS will allow the Earth science community to make a step change in developing new concepts and tools for key answers to scientific and socio-economic questions concerning geo-hazards and geo-resources as well as Earth sciences applications to the environment and human welfare.
The EPOS IT architecture consists of a central system – the place where integration of data, services and software occurs - the so-called Integrated Core Services (ICS). It consists of a portal and catalogue, the latter providing to end-users a 'map' of all EPOS resources (datasets, software, users, computing, equipment/detectors etc.) in compliance with the FAIR principles. ICS is extended to ICS-d (distributed ICS) for certain services (such as visualisation software services or Cloud computing resources) and for specific simulation or analytical processing. ICS also communicates with TCS (Thematic Core Services) which represent European-wide portals to national and local assets, resources and services in the various specific domains (e.g. seismology, volcanology, geodesy) of EPOS. The three layer metadata architecture management model and its implementation in the ICS system catalogue are

[15] www.slideshare.net/BlueBridgeVREs/epos-metadata-catalogue
[16] www.epos-ip.org/

currently built upon an international standard relational data model for storage and interoperability of research information, CERIF (Common European Research Infrastructure Format).

Research data catalogues and data interoperability in life sciences, Rafael Jimenez, ELIXIR[17]

ELIXIR[18] is an intergovernmental organisation that brings together life science resources from across Europe. These resources include databases, software tools, training materials, cloud storage and supercomputers. The goal of ELIXIR is to coordinate these resources so that they form a single infrastructure. The ELIXIR infrastructure makes it easier for scientists to find and share data, exchange expertise, and agree on best practices. ELIXIR services are provided by ELIXIR Nodes.

ELIXIR has several data catalogues for life sciences based on different approaches of data interoperability and federation.

How FAIR is GEOSS, Mattia Santoro, CNR-IIA[19]

GEOSS[20], the Global Earth Observation System of Systems is a set of coordinated, independent Earth observation, information and processing systems that interact and provide access to diverse information for a broad range of users in both public and private sectors. GEOSS links these systems to strengthen the monitoring of the state of the Earth. It facilitates the sharing of environmental data and information collected from the large array of observing systems contributed by countries and organizations within GEO. Further, GEOSS ensures that these data are accessible, of identified quality and provenance, and interoperable to support the development of tools and the delivery of information services. Thus, GEOSS increases our understanding of Earth processes and enhances predictive capabilities that underpin sound decision-making: it provides access to data, information and knowledge to a wide variety of users.

This 'system of systems', through its Common Infrastructure (GCI), proactively links together existing and planned observing systems around the world and support the need for the development of new systems where gaps currently exist. It will promote common technical standards so that data from the thousands of different instruments can be combined into coherent data sets.

However, is GEOSS addressing the FAIR principles?

- Findable: GEOSS has to deal with the large amount of datasets provided by the end systems and to collect metadata and provide effective discoverability. Dealing with such numbers, normally constrained queries commonly match a large number of datatsets. GCI addresses this challenge by returning a smaller and/or an ordered result sets.

[17] www.slideshare.net/BlueBridgeVREs/research-data-catalogues-and-data-interoperability-in-life-sciences
[18] www.elixir-europe.org
[19] www.slideshare.net/BlueBridgeVREs/how-fair-is-geoss
[20] www.earthobservations.org/geoss.php

- Accessible: In GEOSS, main Access-related challenges include: visualisation of data previews and basic transformations (that is to provide users with an easy access to the discovered data along with a set of basic data transformations to make them more easily processed). The DAB + GEOSS Portal access transformation allows to deliver discovered datasets according to a common grid.
- Interoperable: All GEOSS components are bound by the requirements on contributed systems. Therefore, extensive efforts have been invested on interoperability of data and information systems.
- Reusability: In GEOSS, challenges related to Re-usability mainly stem from datasets heterogeneity. In addition, GEOSS needs to address the requirement to support diverse cross-disciplinary applications targeting different communities and user categories which have different needs, as for data discovery and presentation in an informative and significant way.

In the past 10 years GEOSS has developed a truly Global and multidisciplinary System-of Systems. A valuable framework to experiment and learn how to deal with Multi-disciplinary Interoperability challenges. The new GEOSS Portal + DAB platform significantly improved the discoverability and accessibility of shared GEOSS resources, addressing more and more the User requirements. Creating a user oriented open science approach is vital to maximise and incentivise user engagement.

EUDAT-B2FIND: A FAIR-friendly and interdisciplinary data catalogue, Heinrich Widmann, German Climate Computing Center (DKRZ)[21]

EUDAT[22], the European Data Infrastructure, establishes a pan-European e-infrastructure supporting multiple and diverse research communities in their research data management. This Collaborative Data Infrastructure (CDI) is based on the FAIR principles and implements community-driven and generic services to tackle the specific challenges of international and interdisciplinary research data management.
The EUDAT metadata service B2FIND[23] plays a significant role in this context as a repository and a search portal for the diverse metadata collected from heterogeneous sources. For this, EUDAT built up a comprehensive joint metadata catalogue and an open data discovery portal and offer support for new communities interested in publishing their data within EUDAT.
Within the B2FIND ingestion workflow the mapping of the non-uniform, community specific metadata to homogenous structured datasets is the most subtle and challenging task. The homogenisation of the community specific data models and vocabularies enables researchers to search within a comprehensive data catalogue using the powerful B2FIND discovery portal. Furthermore, the service provides transparent access to the scientific data objects through the given references in the metadata.

[21] https://www.slideshare.net/BlueBridgeVREs/eudatb2find-a-fairfriendly-and-interdisciplinary-data-catalogue
[22] www.eudat.eu
[23] http://b2find.eudat.eu/

"An ecosystem to support FAIR Data", Luiz Olavo Bonino, Dutch Techcentre & Vrije University Amsterdam[24]

As one of the organisations present at the Lorentz workshop in January 2014 where the concept of FAIR Data has been created, the Dutch Techcentre for Life Sciences has, since then, worked on a number of solutions to support the adoption and dissemination of the FAIR Data Principles among with:

- BYOD (Bring Your Own Data) to learn how to make data linkable "hands-on" with experts; to create a "telling story" to demonstrate its use and to make FAIR Data at the source;
- FAIRifier that is a FAIRification process consisting of i) retrieving original data; ii) identifying and analysing datasets; iii) defining the semantic model; iv) transforming data; v) assigning a license; vi) defining metadata and vii) deploying FAIR Data resource (data, metadata, license);
- FAIR DATA POINT: A particular class of FAIR Data System that provides access to datasets in a FAIR manner. The datasets can be external or internal to the FAIR Data Point. Also, the source data can be a non-FAIR dataset or a FAIR Data Resource. If the source data is non-FAIR, the FAIR Data Point needs to made the necessary FAIR transformations on the fly.

[24] https://www.slideshare.net/BlueBridgeVREs/an-ecosystem-to-support-fair-data

4. Towards a common EOSC catalogue: workshop take-aways

Once the presentations were delivered, the audience and the speakers were asked to reflect on the following topic and provide input on three questions:
"All the existing infrastructures are called to contribute to the European Open Science Cloud (EOSC). According the EOSC HLEG it is expected to be "a commons based on scientific data" built starting from these infrastructures. In this role EOSC will have to provide capabilities for finding, accessing, interoperating and reusing data."

1. How could the EOSC catalogue facility be implemented? Through a single global data catalogue that gathers the metadata of all the published "data"? By harvesting metadata from the participating infrastructure data catalogues? Or what other model do you envisage as the most appropriate?
2. Should EOSC aim at introducing a single, even if minimal, common metadata format that is used by each infrastructure to publish data outside its boundaries or should we introduce mediators between metadata formats as basic components of the EOSC architecture?
3. Currently each infrastructure has its own publication policy. Should EOSC impose a set of common policies on what, when and under which conditions data can be published in the catalogue?

In this section a summary of the views of the participants are reported:

Question1:
How could the EOSC catalogue facility be implemented? Through a single global data catalogue that gathers the metadata of all the published "data"? By harvesting metadata from the participating infrastructure data catalogues? Or what other model do you envisage as the most appropriate?

- All the contributors agree on the fact that EOSC should offer a data catalogue to its users and that it has necessarily be built as a "Catalogue of catalogues" where existing catalogues can be national, institutional, discipline or project specific ones;
- There are several infrastructures and initiatives that are already making an effort to integrate data and metadata from multiple catalogues. They are adopting different solutions more or less based on shared protocols and standards. The EOSC Catalogue can rely on the work already done by these initiatives.
- It is certainly unrealistic at the moment to assume that the EOSC Catalogue can be built by asking to all the existing component catalogues to adopt common metadata standards and interfaces. Reaching this harmonization requires many changes and years of works from participating actors. This common solution may possibly be reached in the long term when the return of investment of sharing will be well understood. In the meantime, more pragmatic solutions, based on ad-hoc transformations and mediators that do not necessarily require considerable changes

in existing catalogues should be put in place. In parallel, actions can be done to progressively introduce shared guidelines starting from very simple ones. Catalogues presented in the workshop showed few commonalities. Initial guidelines might leverage them.

- Creating a new "catalogue of catalogues" is a solution that presents some risks: a) maintenance of the catalogue, b) creation and update of entries, c) choice of a model to store metadata in an appropriate way d) technical issues: is it really feasible? e) scalability and granularity: How to group and structure the metadata from the various sources?
- The model selected to store metadata should take into account both the capacity of representing "rich" metadata (i.e. metadata related to different concepts, as for instance users, datasets, catalogues, projects, equipment etc.), and the possibility of dealing with semantics in a smart way, (i.e. use multi-domain ontologies, or support the capability of representing mappings among different ontologies).
- It would also be useful to define a common taxonomy to classifiy data (public and open data, private and big data, sensitive data, anonymized personal data).
- One potential disadvantage for a single catalogue of catalogues may be that community specific fields and 'search interfaces' could not be offered. It would be important to identify solutions that enable to refer to the more specific catalogue information when needed.

Question 2:

Should EOSC aim at introducing a single, even if minimal, common metadata format that is used by each infrastructure to publish data outside its boundaries or should we introduce mediators between metadata formats as basic components of the EOSC architecture?

The inputs collected from the contributors are quite diverse and reported below:

- The adoption of a minimal common metadata format with associated protocols can be useful in the case where it is of interest the discovery of the available resources. However, this must be 'extendable' by templates or something similar.
- As findable data is depending on rich use of metadata, a minimal format will not make it easier to find data. Working towards use of semantic metadata description will facilitate easier exchange of metadata between different formats.
- Richest metadata formats can be complex to adopt, but have the advantage to make the data more "usable" by both humans and machines, that through a detailed and rich metadata description can filter, select, process, or even visualise data and data products in an appropriate way.
- With both the solutions proposed some common actions need to be performed:

    - To promote the adoption by Research Infrastructures and e-Infrastructures of already existing metadata standards (e.g. INSPIRE, OGC etc.), protocols and practices.
    - To promote the best practice of publishing metadata in multiple formats thus to match different needs. Such formats include both community specific standards (e.g. Darwin Core), data type specific standards (e.g. ISO 19115), as well as community agnostic / generic Standards (e.g. Dublin Core, Schema.org)
- It is important to reduce the barriers to contribution to such catalogue as far as possible: a model where people provide metadata in the format that work best for them is the solution. Forcing people to provide data in a way decided by externals will reduce adoption.

Question 3:

Currently each infrastructure has its own publication policies. Should EOSC impose a set of common policies on what, when and under which conditions data can be published in the catalogue?

- Common policies, particularly if reinforced by funding policies, can be very helpful. A clear set of guidelines and recommendations for the data providers should be envisaged with regards to the provided metadata and to the underlying data collections.
- The recommendation is that the common policies focus initially on ensuring that the data is FAIR. Disseminating and pushing to create a FAIR culture is the way to go if shared principles and publication policies must be adopted.
- Looking for greater consistency on data licenses would be the next thing to tackle.
- Already quite a lot of work has been carried out on publication policies, so it is recommended to maximise the re-usage of existing results. Indeed, only a balance between top down and bottom up approach, (i.e. the co-development approach) can ensure that solutions are agreed and finally adopted.
- EU-funded research could impose a common policy, but any other research should be able to have its own policy. However, it must be considered that imposing a set of rules might turn out in an action without results.
- Relying on clustering initiatives (e.g. ESFRIs) is probably a good opportunity to be sure that communities are involved and adopting policies.

Annex1

The annex reports the individual contributions of the participants who contributed to the report.

Question 1: How could the EOSC catalogue facility be implemented? Through a single global data catalogue that gathers the metadata of all the published "data"? By harvesting metadata from the participating infrastructure data catalogues? Or what other model do you envisage as the most appropriate?

- **Massimiliano Assante, CNR**: Assuming that each participating infrastructure provides its own data catalogue, the EOSC catalogue facility might be implemented as a Catalogue of Catalogues. It should offer to its users the possibility to transparently querying it also using specific metadata formats (or set of metadata formats) even if not natively supported by the underlying original catalogues. To enable this behavior in a so heterogeneous context as the one delineated by the existing catalogues a mix of technical solutions will have to be supported. These will have to combine harvesting into a central catalogue with distributed search and access facilities according to characteristics and policies of the interfaced catalogues.
- **Daniele Bailo, INGV:** In the framework of the EOSC is of course fundamental to have access to heterogeneous resources in a simple way. With this objective in mind, creating a new catalogue is one of the viable options. However, a serious discussion should be undertaken about the effectivity of this solution in order to match the objective (i.e. facilitate access to heterogeneous EOSC resources). Creating a new "catalogue of catalogues" is a solution that presents some risks or issues: a) maintenance of the catalogue, b) creation and update of entries, c) choice of a model to store metadata in an appropriate way. In this sense, previous to the question "should we create a new catalogue", a harmonisation activity that promotes the adoption of common metadata standards and interfaces, that in turn will enable existing catalogues to expose their metadata in a machine-understandable way, should be carried on. Such an initiative will improve interoperability of system. With this premise, also the adoption and creation of a new catalogue, becomes an action whose risks are mitigated: a) maintenance and b) updates of entries can be done in an automated way by harvesting metadata from participating infrastructure catalogues; c) one of the models now used to store metadata can then be adopted. Such a model should take into account both the capacity of representing "rich" metadata (i.e. metadata related to different concepts, as for instance users, datasets, catalogues, projects, equipment etc.), and the possibility of dealing with semantics in a smart way, (i.e. use multi-domain ontologies, or support the capability of representing mappings among different ontologies).
- **Ramon Codina, Communications Maritime Hub:** It is important first to define a common criteria science taxonomy and classification data (public and open data,

private and big data, sensitive data, anonymized personal data). I propose to include in EOSC a metadata about ecological footprint and biocapacity (see http://data.footprintnetwork.org). Into the EEZ (Economic Exclusive Zone) 200NM we propose to use our international initiative IaaS (Communication Maritime Hub) Connecting the Smart Sea (Oceanography observatories, Oceanographic buoys of EMSO ERIC, and marine rescue), Smart Port (Cruises and Sustainable Shipping and Port Logistics) and People in a Smart Maritime Hub with a very low latency and a coverage mobile broadband into the EEZ, see our propose in https://eu-smartcities.eu/commitment/2621.

- **Harry Lankreijer, ICOS**: Could one single global data catalogue be technically feasible and make data easy accessible. Harvesting metadata from other catalogues seems to be a better solution. Either way, rich metadata is essential.
- **Andrew Treloar, ANDS[25]:** The catalogue facility should bootstrap on existing endeavours. These might be national (such as the Dutch NARCIS), institutional, discipline or project. The catalogue should aim to harvest from these into a single catalogue, and remove any duplications along the way. A model for how to do this at a national scale that could be generalised is http://researchdata.ands.org.au/, run by the Australian National Data Service. The approach that we use and all the source code are freely available for adoption by EOSC if that is useful.
- **Heinrich Widmann, EUDAT:** One single global data catalogue would be the approach as implemented by EUDAT-B2FIND. The advantage is that users must access only one single entry point (interface) to search in a comprehensive and common search space. The disadvantage may be that you cannot offer community specific fields and 'search interfaces' and that you have to homogenize to one common schema. Other issues to be considered with this global approach are scalability and granularity: How to group and structure the metadata from the various sources?

Q2. Should EOSC aim at introducing a single, even if minimal, common metadata format that is used by each infrastructure to publish data outside its boundaries or should we introduce mediators between metadata formats as basic components of the EOSC architecture?

- **Massimiliano Assante, CNR:** The EOSC catalogue facility cannot rely on a single, even if minimal, common metadata format so as to fall under the "Agreement-based" approaches for interoperability. There is a need to guarantee a high level of autonomy among the partaking infrastructures. Thus it is required to use approaches able to

[25] This contribution should be read as a personal view.

isolate the interoperability machinery and implement it in mediators between metadata formats.

- **Daniele Bailo, INGV:** When planning and promoting the adoption of European wide models, rules and standards, it is of great importance to take into account technical and social issues and also to focus on the objectives. The two options proposed in the questions are both interesting according to the goal they want to pursue. The adoption of a minimal common metadata format with associated protocols (for instance Dublin Core and OAI-PMH) can be useful in the case where it is of interest the discovery of the available resources. Then a manual refinement is required if a user wants to access the actual data (or - in general - resources). Richest metadata formats can be complex to adopt, but have the advantage to make the data more "usable" both by humans and machines, that through a detailed an rich metadata description can filter, select, process, or even visualise data and data products in an appropriate way. In order to match both objectives and maximise impact, some key principles might be outlined, for instance:
    - promote the adoption by Research Infrastructures and e-Infrastructure of already existing metadata standards (e.g. INSPIRE, OGC etc.)
    - promote the best practice of publishing metadata both in rich metadata standards (sometimes very community specific) and in generic standards (Dublin core).

  With this premise, the creation of a mediator, which is a task to which much resources should be dedicated, can become simple and - with the adoption of appropriate metadata models - feasible. In any case, building on the experience of Research Infrastructure like EPOS, I think that the EOSC should be used as an opportunity to harmonise data, metadata, best practices and tools. Questions like "should we build a catalogue" or "should we build a mediator" are out of the scope at the moment. I think we should FIRST start from a common basis where all RIs and e-Infrastructures adopt common standards, protocols and practices. With this premise, new scenarios will open up, where anybody (even skilled IT users) could harvest metadata, build their own mediators or applications (even mobile). Likewise, with the above premise, access to resources and building of catalogues will be simpler. The creation of a mediator would be facilitated and even several mediators could be built, according to the needs of specific communities and domains.
- **Ramon Codina, Communications Maritime Hub:** A common metadata format, or common criteria science taxonomy is basic. EOSC should aim at introducing this common criteria science taxonomy to publish data or open data. The EOSC architecture must define an API, and the controller rule must be mandatory in big data and open data. In case to use a sensitive data Binding European Research Council Rules

must be mandatory to all European Research Council members[26]. See an example list of BCR at http://ec.europa.eu/justice/data-protection/international-transfers/binding-corporate-rules/bcr_cooperation/index_en.htm

- **Harry Lankreijer, ICOS:** Today the work done on metadata standards is going towards a certain common minimum. However, findable data is depending on rich use of metadata and thus a minimal format will not make it easier to find data. Working towards use of semantic metadata description will facilitate easier exchange of metadata between different formats.
- **Andrew Treloar[27], ANDS:** The challenge in producing such a catalogue is that the catalogue producers care more about getting the data instead of caring about providing data. In other words, it's important to reduce the barriers to contribution as far as possible. In the early days of ANDS we were able to provide funding to data producers to do things "our way". Once we were no longer providing such funding, many of the feeds dried up. We needed to move to a model where people provided metadata in the format that worked best for them and we took on the work of converting this into what we needed. So, I would argue for the mediator approach.
- **Heinrich Widmann, EUDAT:** I would suggest a minimal, common metadata schema. However, this must be 'extendable' by templates or something similar. "Mediators between metadata formats" sound nice as well, but I have no idea how they should be implemented.

Q3. Currently each infrastructure has its own publication policies. Should EOSC impose a set of common policies on what, when and under which conditions data can be published in the catalogue?

- **Massimiliano Assante, CNR:** Both the EOSC High Level Expert Group report and the successive GO-FAIR reports suggest choosing a lightweight integration at the level of EOSC. This is also confirmed by our experience in dealing with the federation of heterogeneous providers. The engagement rules may progressively become more prescriptive once the EOSC emerges as a useful and operational reality. For example, initial light rules might be limited to imposing the specification for any item in the catalogue of its terms of use and of a persistent identifier while all the other description fields might be optional both in the term of the format and in the used vocabulary.

- **Daniele Bailo, INGV:** Already quite a lot of work has been carried out on publication policies, so we should maximise the re-usage of existing results. Indeed, only a balance between top down and bottom up approach, (i.e. the co-development approach) can

[26] https://erc.europa.eu/about-erc/mission
[27] This contribution should be read as a personal view.

ensure that solutions are agreed and finally adopted. Imposing another set of rules might turn out in an action without results. In Europe, we already have INSPIRE regulations and indications. Creative Commons licenses are often adopted by many RIs. Additionally, for data publication in a commons we have OpenAIRE. Disseminating and pushing to create a FAIR culture is the way to go if shared principles and publication policies must be adopted. Finally, relying on clustering initiatives (e.g. ESFRIs) is probably a good opportunity to be sure that communities are involved and adopting policies.

- **Ramon Codina, Communications Maritime Hub:** It's important the rule of the Data Privacy Officer (DPO) in all European Research Council members if EOSC were to use a model of DaaS (Data as a Service) and compliance with the F.A.I.R. data principles. The rule of the DPO is to ensure to the digital society the best practice about controllers and processors, in a model of DaaS (Data as a Service) and compliance with the F.A.I.R. data principles.
- **Harry Lankreijer, ICOS:** If the aim is to obtain as many data as possible for re-use in the data catalogue, the researcher should be motivated to publish the data. EU-funded research could impose a common policy, but any other research should be able to have its own policy. However, researchers should be motivated to publish by seeing the benefits of it: increased chances for funding. However, this will need also a good system for citation to published and downloaded data.
- **Andrew Treloar[28], ANDS:** Common policies, particularly if reinforced by funding policies, can be very helpful. I would recommend that these policies focus initially on ensuring that the data is FAIR. Looking for greater consistency on data licenses would be the next thing to tackle.
- **Heinrich Widmann, EUDAT:** Yes, at least regarding the provided metadata there should be a clear set of guidelines and recommendations for the data providers. Another thing are the policies of the underlying data collections, e.g. the access permissions of the data resources may differ between the infrastructures - and that's ok, as long this is clearly specified in the metadata (e.g. in a field 'Licences' or 'Rights').

[28] This contribution should be read as a personal view

SPECIAL THANKS TO

Massimiliano Assante, CNR- ISTI
Daniele Bailo, INGV
Luiz Olavo Bonino, Dutch Techcentre & Vrije University Amsterdam
Donatella Castelli, CNR- ISTI & BlueBRIDGE Coordinator
Ramon Codina, Communications Maritime Hub
Rafael Jimenez, ELIXIR
Harry Lankreijer, Lund University
Mattia Santoro, CNR-IIA
Andrew Treolar, ANDS
Heinrich Widmann, German Climate Computing Center (DKRZ)

…and to all the workshop participants

BlueBRIDGE Consortium Members

Coordinator: **Italian National Research Council - CNR** (IT)

Partners: **ERCIM** (FR); **Engineering** (IT); **University of Athens** (GR); **The Food and Agriculture Organisation of the United Nations – FAO** (IT); **The International Council for the Exploration of the Sea – ICES** (DK); **Institut de recherche pour le developpement – IRD** (FR); **Foundation for Research and Technology Hellas – FORTH** (GR); **Trust-IT Services Ltd** (UK); **I2S** (GR); **CITE** (GR), **Collecte Localisation Satellites SA – CLS** (FR); **GRID-Arendal** (NO); **Pole Mer Bretagne Atlantique** (FR).

Find out more about BlueBRIDGE www.bluebridge-vres.eu

Follow us on Twitter @BlueBridgeVREs

BlueBRIDGE receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 675680