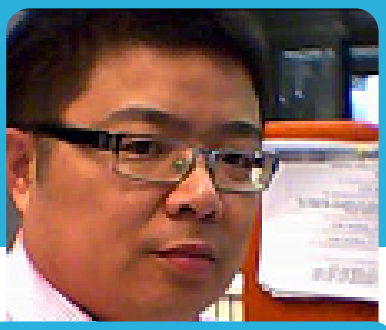
ENVRI-FAIR NEWSLETTER
MAY 2021

DANIELE BAILO
ENVRI
SERVICE CATALOGUE



ZHIMING ZHAO
ENVRI
KNOWLEDGE BASE

ARI ASMI
ENVRI USE CASE

THE THREE COMPONENTS TO ENVRI-HUB SUCCESS

In this interview, Daniele Bailo, Zhiming Zhao, and Ari Asmi explain to you the key components forming the ENVRI-hub

Daniele, can you explain a little bit more in detail what are these three components crucial to ENVRI-hub?

Daniele: These are three distinct but complementary elements used as containers to convey the provision of heterogeneous datasets, services, and software in ENVRI.

The **ENVRI catalogue** concept is about creating a **catalogue of services** now implemented and operating at the research infrastructure level. By accessing or harvesting such a catalogue, an external initiative like the European Open Science Cloud can obtain machine-understandable and machine-readable information to interoperate with the FAIR services from the sub-domain research infrastructures.

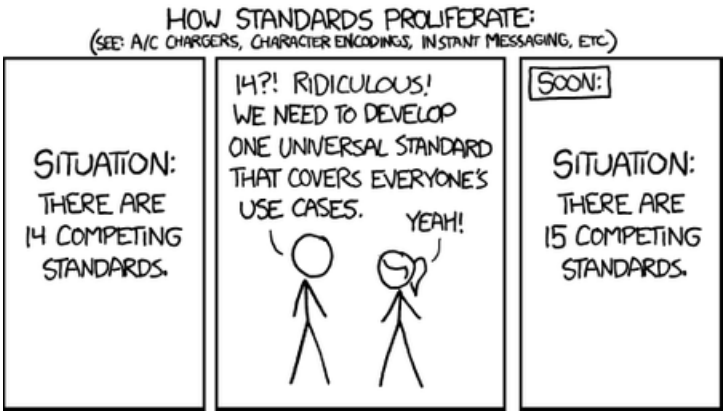
The **ENVRI Knowledge base** is a container of metadata describing datasets from the communities willing to provide these. Technically speaking, the metadata can be harvested by the knowledge base from the sub-domain research infrastructures that implement a common metadata standard. This work is likely to improve the FAIRness level of all research infrastructures in ENVRI considerably.

The **Science Use Cases** building block is where the missing elements, i.e., the processing and analysis of data, are considered within ENVRI. In particular, the Science Use cases will rely on Jupyter notebooks.

ENVRI SERVICE CATALOGUE

I suppose you have to define some standard metadata ontology to describe the catalogue of services and simplify the search of each specific service. Can you tell us a bit more about the work related to this?

Daniele: An attempt to define the definitive and comprehensive metadata ontology usually ends up in a well-known joke. The joke is about a scientist named Bob. A colleague asks Bob to make an order among fourteen non-interoperable metadata standards. And what Bob does? He builds a new standard encompassing them all. After few months, the colleague comes back again and says: “Hi Bob.... We have a problem: we have fifteen non-interoperable metadata standards....”.



"How standards proliferate" comic: https://xkcd.com/927/

The moral of the story is: **it is hard to find a final, rich and good metadata ontology that works for all cases, in all cases**. However, by making some reasonable compromises, it is possible to step forward and achieve an improved FAIRness. In the ENVRI-FAIR case, the proposal to adopt DCAT-AP v2 and its extension came almost naturally from different actors. Note that we consider this standard as a way to expose metadata information. Specific communities will, of course, keep on describing and storing metadata and ontologies with the format they know and are used to work with. CERIF, for instance, seems to be a good candidate for reflecting the complexity of rich metadata.

The ENVRI service catalogue will be ultimately integrated into the European Open Science Cloud. The vision of EOSC is to be an environment for hosting and processing all kinds of research data. How crucial is the ENVRI contribution to the EOSC?

Daniele: The driving idea here is that ENVRI-hub is making the ENVRI community "EOSC prepared". It means that once the services are catalogued in the ENVRI catalogue of services, they are more ready for the EOSC. This way, the research infrastructure services will reach more users without exporting or dumping their main assets, i.e., the datasets. Instead, research infrastructures will expose the metadata harvested and reused to make their assets FAIRly accessible.

I would like to make a point relevant to the entire EOSC setup: the research infrastructures, people behind them, and the related initiatives, like ENVRI-FAIR, are the ones making the EOSC concept possible. Without research infrastructures that guarantee the long-term sustainability of data services, no integration and analysis are possible. In this sense, I envisage an improved involvement of the research infrastructures at the governance and financial level in this European Wide initiative.

When can we expect the catalogue will be implemented and accessible by users?

Daniele: We envisage a first beta version prototype of Service Catalogue to be ready for testing within the ENVRI community by the end of 2021. Depending on the results, the complexity of the task, and community reactions, a more robust version of the catalogue may be available during the second half of 2022.

ENVRI KNOWLEDGE BASE

Zhiming, ENVRI-FAIR is already the third project supporting the collaboration among the research infrastructures. I suppose there is a lot of know-how on technical practices. We can offer solutions to common problems related to interoperability challenges among the environmental research infrastructures. Is there a place where we can search for all these technical practices?

Zhiming: **Yes, we can now search such information from a centralized search engine, namely the ENVRI knowledge base search engine.** The search engine indexes the technical knowledge in the knowledge base collected from different research infrastructures during their development activities, FAIRness assessment results of the research infrastructures, and online information from the ENVRI community, including documents from their websites and data sets services, and APIs from their online catalogues. The search engine is developed based on the open semantic search framework. It periodically crawls and indexes online information from the ENVRI online sources and allows human operators to adapt the crawling policies and curate the indexes.

I believe the primary users of the knowledge base are the research infrastructures themselves. Do you foresee the base could be helpful for other users as well?

Zhiming: Yes, you are right. The current primary users are the developers and data managers from the research infrastructures in ENVRI. These users contribute to the content and join the co-development of the knowledge base. However, the functionality offered by the knowledge base, specifically the search engine, can certainly serve a much broader scope of users, such as domain

scientists, newcomers to the ENVRI community, policymakers, and students. We can see those needs from the discussion when we demonstrate the current version during the last ENVRI week.

Moreover, we think the search engine will also support commercial users eventually, such as airlines, mining, or shipping companies. Commercial users need the ENVRI data or services to realize their business values, often with customized agreements for data access and usage. In this way, they also contribute to the sustainability of the ENVRI services.

When do you expect the knowledge base will be fully functional and populated with the content?

Zhiming: The initial version has already been online since 2020 (http://search.envri.eu). We follow a co-design and iterative development strategy; we aim to update the knowledge content and search functionality continuously. In addition, we plan to engage the knowledge curators from research infrastructures to customize the indexing pipeline, control the quality, recommend additional knowledge sources, and improve the efficiency of the knowledge base management.

ENVRI USE CASES

Ari, one of the main goals of the European Open Science Cloud is to offer services that are interoperable, which will eventually foster interdisciplinary research. Can you give us some examples of challenges that require a cross-border research collaboration?

Ari: If by borders, you mean the disciplinary ones, there are many such challenges. Most of the climate change and ecosystem services-related problems are by their nature cross-disciplinary, particularly when considering the wide range of processes from physicochemical reactions all the way to societal and ecosystem responses. These come even more important when trying to answer more abstract societal goals, such as long-term availability of food and water or inequalities related to ecosystem changes. Overall, most truly humanity-encompassing challenges always require a holistic approach and need multidisciplinary science to answer them. That goes even further than (already very wide) disciplinary fields of environmental research infrastructures.

One good example that demonstrated this is the COVID-19 crisis — our sister cluster, EOSC-LIFE, delivered a significant part of the EU's research & development response to the problem. However, to provide such a response, they also needed additional services from other research infrastructure clusters...Simply, none of the grand challenges can be solved without interdisciplinary research.

Is ENVRI-FAIR going to demonstrate how to leverage the services offered by environmental research infrastructures for interdisciplinary research?

Ari: Some of the ENVRI-FAIR subdomain work packages have a work plan with concrete examples for their subdomain products. Additionally, **the ENVRI-hub concept includes a Jupyter notebook environment with practical examples of real-world scientific use cases using the hub interfaces and methods.** Jupyter notebooks are basically like a journal article that also contains the actual research steps.

I believe these practical examples are perhaps the most direct and efficient way to demonstrate and guide how scientists can build their workflows and tools around the research infrastructure services.

Outside of the project itself, the recently started EOSC FUTURE project will include some genuinely cross-cluster examples with ENVRI participation - such as work on invasive species analysis and a dashboard for the state of the environment. These are all just demonstrations, of course. They are more intended as examples that user communities could use to build their own research environments for their own particular needs. We hope we have many of such user communities in the future.