EOSC-Pillar

Coordination and Harmonisation of National &Thematic Initiatives to support EOSC

Second Annual Report

Building upon the Foundations

EOSC-Pillar Second Annual Report

The EOSC-Pillar Annual Report is a publication of the EOSC-Pillar project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 857650 and is produced to showcase the major results and achievements of the project, collaborations ongoing with other initiatives and updates for the wider community.

Table of contents

Foreword

Marialuisa Lavitrano

EOSC Association Director

This first year of my service as one of the Directors of the EOSC Association has been a bit of a rollercoaster: It had the thrill of new beginnings, and the same uncertainty. Lots of things to initiate, problems to solve, processes that had to be invented, much enthusiasm, and yet more work. Nevertheless, after a year of hard work for us directors - and so many other people - we can't help but be satisfied of the direction we are impressing to the EOSC Association and EOSC in general. It has been an intense year, but we laid down a path towards a future where a new, more participative and efficient science can be glimpsed in the distance – but not too far away. Meanwhile, a lot of things are moving in the Member States, despite all the difficulties of the pandemic or maybe also because of that, as while some doors may have been closed, at least for a bit, the funding mobilised by the EC recovery plans are possibly opening new ones.

National and regional Open Science initiatives are emerging and growing in many countries while Open Science and data sharing in general catch the attention of other societal strata than just researchers, such as policy makers, public administrations and citizens. Because, let's face it, the pandemic taught us, the hard way, why it is important to share data and to make them FAIR by design. This newly found interest must be nurtured and translated into action, because today's pandemic will be by no means the last challenge we face together, and harnessing science will be the best way to respond to critical societal challenges: one above all, since COP26 just ended, and not with all the answers I am afraid, climate change. Data science will have a key role here, and while no one is going to solve such big problems alone, and a pan-European and even wider approach is mandatory, it is important that Countries and Regions are all involved and do their part. To succeed, EOSC and Open Science in general must be both coordinated and widespread, one without the other won't be enough. That's why I look at regional coordination projects like EOSC-Pillar as key initiatives to really turn the EOSC in reality.

National and regional coordination is the baseline for the success and sustainability of EOSC in the long term, especially when we are addressing not just the relatively few excellences in the Research Infrastructures, with their abundance or resources, but also the long tail of science, with its richness and diversity - but also fragmentation, as well as all these other stakeholders that are far less reachable from a top-down, European point of view. Reaching out to researchers, policy makers, civil servants and even citizens, involving and engaging them with Open Science would be all but impossible without national commitment; but above all coordination among all these initiatives and communities is key if we want to build one EOSC and not a collection of its instances and different interpretations. This is true for the technical level, where we need to achieve full interoperability between many different layers and components, but even more so for the policy and cultural point of view, that are even more important if we really are to ensure that Open Science practices become the 'new normal' as it is boldly set out in the EOSC Strategic Research and Innovation Agenda (SRIA). That's the real value of EOSC-Pillar ad its fellow regional projects and I expect that you will continue along this path in the upcoming year.

Interviews with EOSC Association Directors

€900m German Open Science Initiative NFDI Progresses as EOSC Association Gears Up

As part of the efforts of EOSC-Pillar countries to coordinate open science initiatives, Germany's flagship National Research Data Infrastructure (NFDI) has already engaged its first research consortia, just as the EOSC Association kick-starts its activities.

"In 2020, the first nine consortia of research communities started their work. In total, we expect 25 to 30 consortia to receive funding from the NFDI", said EOSC Association Board Member Klaus Tochtermann (ZBW), adding, *"The NFDI is intended to become a federated research data infrastructure at the national level and is thus the most important anchor for EOSC in Germany."*

Open Science has made some impressive moves forward across EU countries in 2020, and Germany was no exception. In an exclusive interview with EOSC-Pillar, one of the three newly-elected directors of the EOSC Association coming from the EOSC-Pillar countries discussed some of the latest exciting developments of the National Research Data Infrastructure (NFDI) in the context of the European Open Science Cloud (EOSC).

Consolidating National Efforts Through NFDI

The main step taken in Germany towards the promotion and adoption of Open Science principles in the context of EOSC was the creation of the National Research Data Infrastructure (NFDI), launched in 2020 with a budget of €900 Million for the next 10 years, and full support at the institutional level.

The NFDI will consist of a number of so-called consortia, deriving from the formation of scientific communities and service providers around specific subareas of research, defined thematically, methodologically, by the objects or by subject groups.

Each consortium will then develop and offer a service portfolio for research data management for its subarea. The proposal template included a chapter requiring consortia to outline how they will connect with international developments, such as EOSC, and how they will ensure compliance with the FAIR principles.

Tochtermann stressed the importance for national Open Science initiatives to take an active role in shaping the EOSC already from this stage, making sure that it is compatible with multiple areas of research across countries.

EOSC-Pillar has already started engaging with NFDI and can serve as a blueprint for how to connect national infrastructure initiatives with EOSC.

The EOSC Association in Germany

National landscapes are crucial for the success of the EOSC, the virtual environment where researchers will have access to open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines by federating existing data infrastructures.

The progress of the newly-established EOSC Association (and of GAIA-X) is being closely followed via exchanges between the relevant actors in Germany, including the Ministry of Education and Research (BMBF), the Council for Information Infrastructures (RfII), the NFDI and the Alliance of Science Organisations.

In addition, institutional contact with the German EOSC Association Director Klaus Tochtermann is close, so it can be assumed that the work of the EOSC Association is heard and can have a relevant impact.

"The many German members of the EOSC Association cover all regions well. Thus, they can serve as a local contact point for other institutions in their region. Another model is known from the Netherlands, where a national Open Science platform has been set up, primarily as a hub for Open Science, but also serving EOSC. The two examples show that we need such contact points in principle. Whether they are organised decentrally as in Germany or centrally as in the Netherlands is irrelevant. In any case, it is important to draw even more attention to the benefits and advantages of an EOSC in lectures, seminars, etc," Tochtermann said.

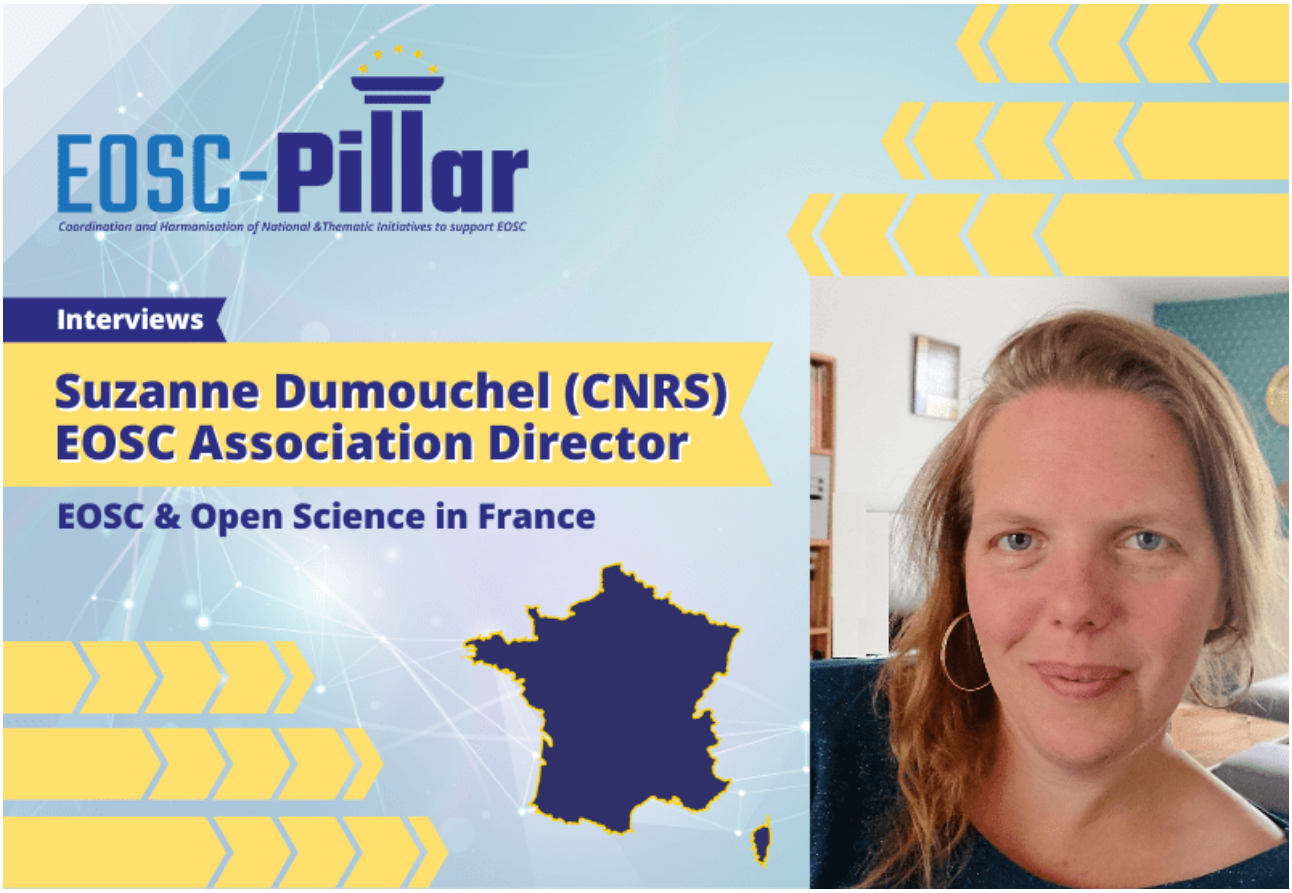
A key issue he identified is that as of the moment, far too few researchers still know far too little about EOSC, as also emerged from the EOSC-Pillar National Initiatives Survey last year.

One of the first tasks of the EOSC Association through 2021, in collaboration with institutions and projects such as EOSC-Pillar, is to bridge this gap with the wider research community and further involve national research organisations in the development and implementation of the European Open Science Cloud.

In conclusion, Tocthermann agreed that a major factor for the recent fast enhancements of Open Science initiatives in Europe was undoubtedly research related to the COVID-19 pandemic, which has once again demonstrated the advantages of OS.

He said, *"It was only through open international collaboration that such rapid successes in genome sequencing, for example, were possible. At the same time, the importance of science communication and the interest in the influence of science on political decisions has greatly increased, as everyone is affected by the lockdowns, for example. Along with this, the importance of the traceability and reproducibility of research results has gained weight, and the transparency associated with Open Science has become more important."*

From Policy to Practice: France CNRS Launches Open Data Research Directorate and Participation in EOSC Association Thrives



EOSC & Open Science in France: An interview with Suzanne Dumouchel

The latest policy efforts related to the European Open Science Cloud have brought many key actors in Europe to the forefront of Open Science initiatives aimed at encouraging OS practices and culture across disciplines. One such example is the recent creation of the "Direction des Données Ouvertes de la Recherche" (DDOR) or Open Data Research Directorate at CNRS in France.

We have explored some of the main updates in the French landscape with newly-elected Director of the EOSC Association Suzanne Dumouchel (CNRS, Huma-Num), Partnership Coordinator of OPERAS.

"In the context of the development of Open Science, French institutions have reacted quickly by creating Open Science Officer positions within their institutions, facilitating collaboration between organisations on infrastructure sharing, supporting the policy of open access publishing and offering numerous training actions. The CNRS, for example, has published its Research Data plan in this perspective, which has led to the creation of a new department, the DDOR, "Direction des Données Ouvertes de la Recherche" or Open Data Research Directorate. In addition, French infrastructures have published a number of reports to prepare and make possible the implementation of their services in the EOSC."

Open Science Moves Forward in France

Among other purposes, the EOSC Association provides a single voice for advocacy and representation for the broader EOSC stakeholder community. As the EOSC-Pillar project coordinates harmonisation and cooperation across Austria, Belgium, France, Germany, and Italy, hearing directly from representatives of the Association will help us understand what is happening across the different Member States in terms of Open Science and EOSC readiness.

Dumouchel pointed out many initiatives among French scientific organisations. For example, the actions undertaken by the French communities within international initiatives such as RDA, where the French National Node has implemented a WG on the certification of data repositories and services. In addition, the training component has

been enriched by actions implemented by URFISTs on Open Science issues.

Since the end of 2018, the Committee for Open Science, implemented by MESRI, has been making a major contribution to structuring the French community in Open Science policy and the dissemination and implementation of its principles.

A similar structure will soon see the light of day in France, the Committee for Digital Infrastructures, which also contributes to structuring the French community.

Dumouchel also added that the establishment of the EOSC Association brought "*an acceleration of the implementation of an Open Science policy within the different organisations, the willingness to work together and to structure the French participation in EOSC*."

In the future, this will especially translate into the creation of an "EOSC College" in which French EOSC actors will be able to work together on a variety of EOSC-related issues, including services, PID, partnerships, infrastructures, and repositories.

Guidelines and Collaboration Between Local and International Actors

As noted by Suzanne Dumouchel, the development of EOSC cannot be done and achieved properly without involving national OS initiatives.

This means that these initiatives are necessary to understand the needs and challenges of national scientific communities, but also to identify tools and services that can contribute to EOSC. Since EOSC is the aggregation and harmonisation of national Open Science activities, failing to strengthen these national links will mean that we won't be able to achieve EOSC. The major difficulty, in this case, is how to harmonise national Open Science practices without minimising the variety that reflects the richness of EOSC.

This is the main difficulty of the different levels that interact before the EOSC: local, regional, national and then European, including transnational initiatives such as EOSC-Pillar. This must be built by a strong willingness at each level to maintain dialogue, relying on each other's expertise.

"*The local and regional levels are [...] essential to involve the scientific communities and to get to know them. At each level, there is a certain vision of Open Science which is also based on a different reality, so national institutions must help local organisations to connect to the EOSC but the EOSC must also take into account local feedback for its implementation. And the stakes are high. The presence of the university network at this stage is fundamental: in charge in general of the buildings, it is at their level (in addition to the national level) that internal storage will be managed, or even in the form of mesocentres.*"

EOSC-Pillar brings together a large number of French organisations, contributing to strengthening the links between them, as well as to the setup of specific collaborations over the longer term. Dumouchel also noted the role of these actors in the involvement of relevant national organisations in the various missions they have to carry out.

Open Science guidelines for providers are not built ex nihilo, but will be a result of this long-term collaboration in European projects. There is a permanent dialogue at the national and European level on this topic, to ensure that these proposals and rules are realistic and shared.

Open Science and Global Challenges

According to Dumouchel, the COVID-19 pandemic did not have a particularly strong impact on the developments described above, as activities were already underway. Nevertheless, it did illustrate all the interest, the necessity but also the complexity of implementing Open Science policy on such sensitive data.

"*This COVID crisis also brought a giant step forward for citizen science, which was more of a theoretical idea and is now practised everywhere. The pandemic has highlighted all the interest in publishing data while illustrating the difficulties on issues of research results.*"

The EOSC Association Director also noted that the Calls for Proposals of the National Fund for Open Science are contributing to accelerating the development of OS rules as well as their assimilation and application by researchers. In fact, at present, the evaluation of researchers itself measures their investment in Open Science policy (in terms of data sharing, chosen modes of publication, etc.).

EOSC-Pillar joins Suzanne Dumouchel and the EOSC Association in supporting the EOSC Community and bringing together all relevant stakeholders to co-design and deploy a European Research Data Commons, where data are findable, accessible, interoperable and reusable (FAIR), and also as open as possible.

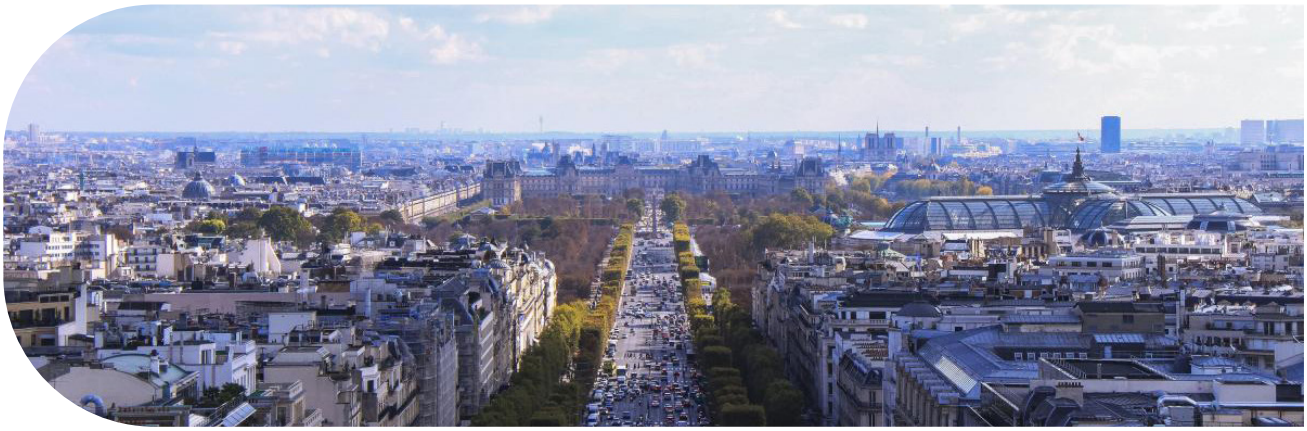
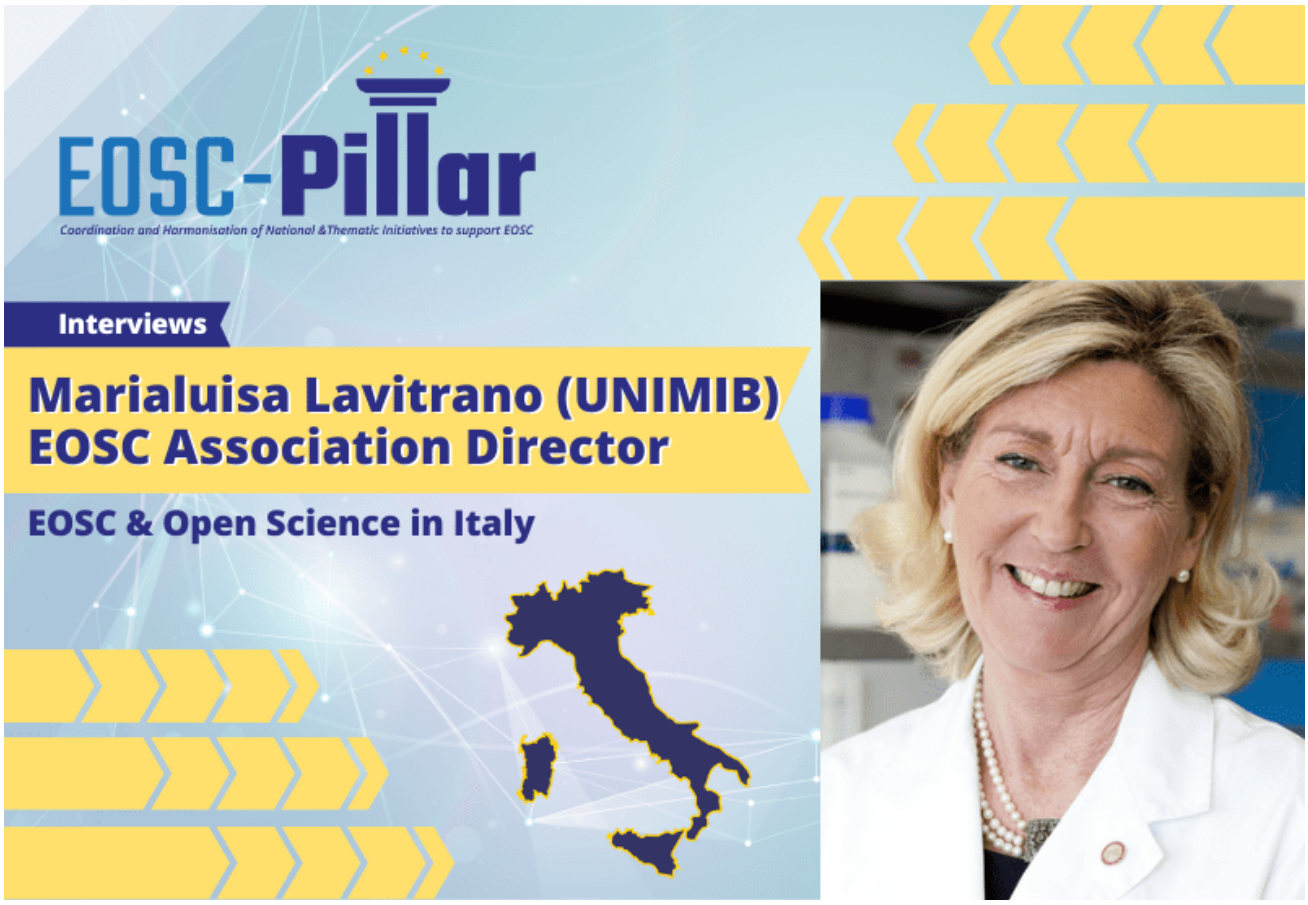


Photo by Pedro Gandra on Unsplash

Turning Points for Open Science in Italy: National OS Plan Advances with ICDI



EOSC & Open Science in Italy: An interview with Marialuisa Lavitrano

2020 was a critical turning point in a wide range of sectors across Europe, and Open Science was no exception. The newly-elected EOSC Association Director Marialuisa Lavitrano (UNIMIB) offered some precious insights on the latest developments in the OS landscape in Italy, such as the evolution of ICDI (Italian Computing and Data Infrastructure) as an actor contributing to fundamental processes for OS.

"[In 2020] we set the basis of what will become the legal entity to drive OS in Italy and represent our country in the EOSC Association: [ICDI] had started the year before as a very lightweight and informal forum where some OS enthusiasts started discussing common issues and ideas, but it was with 2020 that we gained momentum and became really representative of the research and academic scenario."

ICDI can now count on the collaboration of a growing number of universities, many of which are eager to engage with the wider EOSC environment, as shown by the many Italian universities that have applied as members of the EOSC Association.

"2020 was also the moment when we stopped just discussing things and started doing them: we set up shadow working groups to support the work of our national representatives in the EOSC Executive Board ones, and we created three very ambitious and still very practical task forces to support the creation of a national Competence Centre for Open Science, to create a National Federated Cloud dedicated to research and to support the sharing of clinical data - an especially relevant topic these days."

Experts from ICDI have also contributed substantially to drafting the upcoming National Open Science Plan, which will likely play a key role to advance OS in Italy.

EOSC as a Framework for Tackling Global Challenges

Lavitrano stated that EOSC is one of the most ambitious actions ever taken to implement Open Science, and it can build on Europe's unique position in terms of collaboration.

The (relatively) homogeneous legal, regulatory and cultural framework that European countries have in place can be an advantage when addressing a variety of international challenges, which need a significant paradigm shift such as the one required by the Open Science vision.

This was particularly true when the scientific community had to face the COVID-19 pandemic, which according to Lavitrano, *"has dramatically highlighted our lack of prepa-*

ration in a moment where timely data sharing could have literally saved lives."

At the same time, it also spectacularly demonstrated the importance of open data sharing and the need to make it a reality in time for the next big challenge. Data also needs to be FAIR-by-design when such global issues come up because it would be much more complicated to FAIRify data afterwards, while researchers are busy dealing with these major challenges.

As Lavitrano points out, this pandemic might even be considered a fire drill, when compared to climate change, but "*the good news is that Open Science is here to help, so let's make it happen*."

EOSC as an enabler for Open Science in Italy

The past few months have been extremely lively for the OS community in Italy, leading up to the establishment of the EOSC Association and following its first official steps.

"*As the Association became a reality, ICDI received many requests for information and participation from institutions that were not in the OS loop or only marginally by that time. This has led to much better sharing of OS goals and increased interest in joining the soon-to-be legal organisation, but also, at a more practical level, in participating in task forces.*"

The EOSC Association has also increased the level of interest and attention ICDI is receiving from the Ministry of University and Research and from several Academic and Research organisations, which is fundamental in order to think big and start a movement in Italy.

"*A lot of what happens in terms of strategic priorities, enabling infrastructures and funding, happens at the national level, and National Initiatives are essential to leveraging these aspects while gathering a large and active community of Open Science 'believers'*".

It is fundamental that the national and transnational levels are aligned and intertwined to really make things happen, especially at the current stage when so many aspects still need to be defined.

There are no specific national guidelines yet, but with the upcoming approval of the national Open Science Plan along with the creation of a national service catalogue in ICDI (in collaboration with EOSC-Pillar), there will soon be news in this respect.

An interesting aspect of the Italian landscape, according to Lavitrano, is that "in Italy, Open Science was born with a bottom-up initiative and in this spirit, we are doing a lot of community engagement," which is partly compensating for the temporary lack of a national system of incentives.

EOSC-Pillar and ICDI

ICDI also plays a relevant role in relation to EOSC-Pillar, considering that several ICDI members are project partners as well. In particular, GARR, the coordinator of EOSC-Pillar, is also the convenor of the national initiative and represents ICDI as a mandated organisation in the EOSC Association, pending the establishment of the new legal body.

We can say that the initial concept for this project came from the discussion between the Italian and French initiatives, as well as from the recognition needed to encourage the growth of national initiatives and support their role as one of the key pillars for EOSC's success.

There are regular update sessions during ICDI meetings, where the latest developments of the EOSC-Pillar project and the opportunities for the community are discussed. ICDI members are also committed to providing input to EOSC-Pillar activities when this is required, resulting in a fruitful exchange of ideas for both parties.

"*ICDI is a community and therefore can reach out where the Association alone cannot. Many of the members are at the forefront of Open Science in our country, with decades of experience in the different aspects of data sharing, and their advice is highly sought after by research and academic institutions that are just starting to develop their own OS strategy. And this is just the beginning because once ICDI's national Competence Centre is fully operational we will be able to engage many more entities.*"

The plan to create a national service registry and federated data repositories across the country also goes in the same direction. National initiatives will be able to act as a multiplier for EOSC. Last but not least, Lavitrano is confident that a solid backing from ministries and national institutions will help ICDI influence the strategic planning for multidisciplinary research in the future.

Photo by Nicolò Salinetti on Unsplash

Policy and Business Models Work Update

Sara Di Giorgio & Federica Tanlongo, GARR

EOSC-Pillar through WP4 is moving from national initiatives to a viable system of trans-national and ultimately pan-European services. An extensive study on common policies and legal framework was carried out to achieve a common understanding of existing regulations on IPR, Open Data and data protection regulations, also taking into consideration legal and organisational aspects of services delivery in a federated environment, across the five Member States covered by EOSC-Pillar. The study, presented in the Report 4.1 'Legal and Policy framework and federation blueprint', maps differences and gaps in the national regulations, identifies obstacles and fragmentations, and proposes guidelines & a practical checklist to guide researchers and service providers in publishing their data and federating their services in the context of the EOSC cross-border ecosystem. Moreover, recommendations for policy makers were elaborated to highlight issues where an harmonisation effort is needed. A 'second edition' of this document, planned for end of July 2021, will alsol include a comparison of legal and policy topics of our study to RoP and SRIA and perform a more in-depth study into service delivery and service level management.

EOSC Pillar has investigated characteristics of each country in order to consolidate National Initiatives, as strategic nodes for on-boarding services in the EOSC general Catalogue and for animating and engaging with users communities. On the basis of the findings of the WP3 survey, WP4 conducted a qualitative consultation with some selected National Initiatives with the aim of identifying peculiarities and highlighting the points of strength and the areas of improvement. This activity will lead us on tailoring our support offer for the definition of policies and governance, to help them achieve better representativeness in the National and European panorama and in drafting a roadmap that will present guidelines and actions to be performed to consolidate National Initiatives, that will be delivered in the second period.

The consultation with National Initiatives also allowed us to collect important feedback regarding business models and the gaps and opportunities that need to be addressed for the development of EOSC. A Webinar on business and procurement models being co-organised with NI4OS and involving representatives of National agencies, RIs and international organisations such as CODATA and RDA enriched the research WP4 is leading. The output of these activities will be the provision of science-driven recommendations and guidelines for sustainable business models and for lowering barriers to their adoption in an Open Science environment.

EOSC-Pillar through WP4 greatly contributes to promote the cross project collaboration among the INFRAEOSC-05-2018-2019 call (Call5), i.e., *EOSC-Synergy*[1], *EOSC-Pillar*[2], *EOSC Nordic*[3], *NI4OS-Europe*[4], *ExPaNDS*[5], *FAIRsFAIR*[6], and *EOSCSecretariat.eu*[7] An important outcome is the Proposal for Living Indicators to Monitor MS Progresses towards EOSC Readiness, from the Landscape Task Force coordinated by EOSC-Pillar's representative. Based on the Proposal, a Dashboard platform has been developed, thanks to EOSC Secretariat Co-creation funds. It's a PoC that will allow us to better define the workflows and the processes for practically implementing the active monitoring, reporting and analysis of progress against the indicators, including user guidelines and flexible solutions to update and develop new indicators. The PoC methodology and first results are being shared with the new EOSC-Future project and with the EOSC Steering board working group on EOSC readiness indicators, which will build on them to design the future monitoring framework for the EOSC scenario across the EU countries, in order to provide decision makers with educated information for their decisions and ultimately help turning EOSC into a reality in the different regions.

1 https://www.eosc-synergy.eu/
2 https://www.eosc-pillar.eu/
3 https://www.eosc-nordic.eu/
4 https://ni4os.eu/
5 https://expands.eu/
6 https://www.fairsfair.eu/
7 https://www.eoscsecretariat.eu/

National Catalogues and Onboarding

Leonardo Candela CNR & Luciano Gaido, INFN

The main activities concerning national catalogues and onboarding are:

1. A survey about the existence and needs for national catalogues in the EOSC-Pillar countries
2. Definition of a model and prototype for a national catalogue
3. Definition of guidelines to operate a national/thematic/regional catalogue

Existence and Need for National Catalogues in the EOSC-Pillar Countries

To shape the EOSC-Pillar activities about resource catalogues and the support EOSC-Pillar was going to provide, the two main drivers were the characteristics as well as the onboarding procedures of the existing EOSC catalogue and the demand for national catalogues in the five countries participating to the project.

Understanding the EOSC catalogue and its onboarding procedures was quite easy because some EOSC-Pillar partners are directly involved in the projects developing and operating it (namely EOSC-Enhance and EOSC-hub). In addition to this, the activities of the Service Onboarding Task Force, set up among the regional projects (funded in the INFRAEOSC-05 call)[8], were structured in a way to constantly involve representatives of the two above-mentioned projects.

Concerning the other driver, information about existing or planned national catalogues was completely missing. For this reason, a survey was conducted to collect this information at the beginning of Y2.

The information collected showed that things are quite different in the EOSC-Pillar countries. However, it has been highlighted that many national user communities, usually participating in international collaborations and initiatives, have (or plan having) a reference thematic catalogue.

The main outcome of the survey is that in all countries the importance of a national catalogue is recognized but there is no consensus on what a catalogue should provide. In some countries some national catalogues already exist (such as CatOpidor in France, the Ministry's RIs catalogue in Austria or the Open Science catalogue in Belgium) but they have a different purpose and implementation compared to the EOSC Catalogue. Germany and Italy foresee the need for a national service catalogue but its features, scope and goal, as well as the initiative in charge of operating it has still to be defined.

Definition of a Model and Prototype for a National Catalogue

Independently of the specific plans that resulted from the survey, EOSC-Pillar envisaged a key role for both national and thematic catalogues in the overall EOSC onboarding process. Such catalogues are closer to their designated community than the overall EOSC catalogue and have more possibilities to be fully uptaken and developed by the community itself because the process is driven by the target community. Building upon this understanding the project developed a prototype of a national service registry conceived to be interoperable with the EOSC Catalogue and instantiated it by a proof of concept for the italian community[9].

The model underlying this catalogue is innovative. In fact, the catalogue is provided by a web-based working environment promoting a collaborative development of the catalogue itself. The web-based environment offers a networking area where users get informed whenever a new service is onboarded and can easily exchange ideas, requirements and experiences about the onboarded services with the providers as well as with the overall community.

The catalogue was designed to be customizable with respect to the items that can be published by it and the metadata characterising them, e.g. it is possible to configure it to make it possible to register many typologies of services or resources and include in their profiles any information the community is willing to collect. In order to favour the interoperability with the overall EOSC catalogue, it was decided to configure it to support the publishing of Resource Providers and Resources characterised by metadata conforming to the profiles promoted by the EOSC Catalogue[10]. Thanks to this decision the exchange of information between the two catalogues will be simplified (no mapping or transformation is needed).

The overall management of the catalogue (and the accompanying working environment) is in the control of the designated community. In fact, it is up to the community at large to define the objective and scope of the specific catalogue, the policies governing its development and to put in place the actions needed to achieve all of this by simply using the deployed service (e.g. document the decisions by a Wiki, manage membership and roles, approve or reject or update provider and resource information).

Guidelines to Operate a National/Thematic Regional Catalogue

The EOSC catalogue[11] is in operation since a couple of years now and its onboarding procedures, although still evolving to address the evolution of the rules of participa-

8 https://www.eoscsecretariat.eu/communities/eosc-regional-projects
9 EOSC_Pillar IT Service Registry https://eosc-pillar.d4science.org/web/eoscpillaritserviceregistry
10 https://eosc-portal.eu/providers-documentation
11 https://marketplace.eosc-portal.eu/

tion and inclusion criteria and the improvement to the catalogue itself, are well defined and consolidated.

In order to support the national or thematic initiatives willing to operate a national/regional or thematic catalogue which has to be made interoperable with the EOSC catalogue, a document with specific guidelines[12] has been prepared and its first draft has been delivered in June 2021.

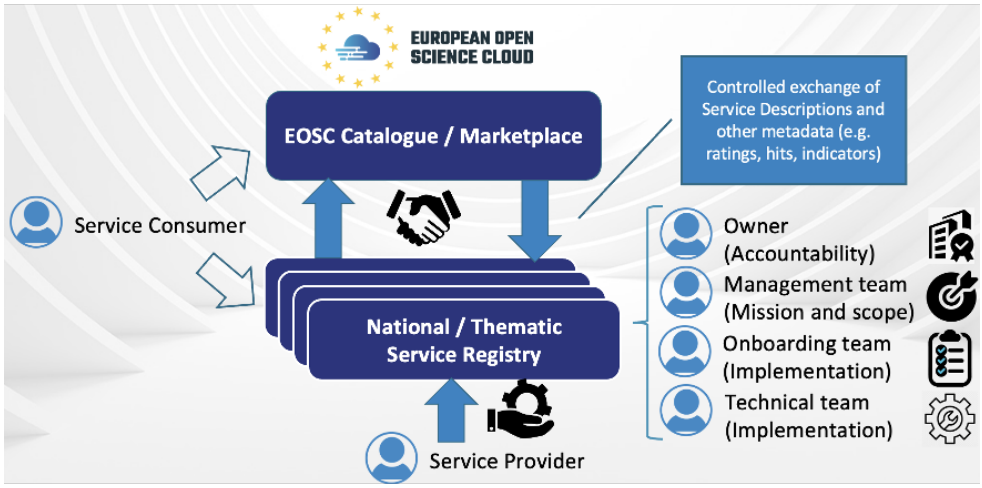
It is intended to provide general and practical help to set up the technical and operational procedures which obviously should be adapted to the specific context of the national/thematic catalogue.

In addition to the general activities needed to operate a catalogue, the specific actions to achieve the interoperability with the EOSC catalogue are described, not only at technical but also at the policy/organizational level, among which one of the most important one is the establishment of an agreement between the parties operating the two catalogues.

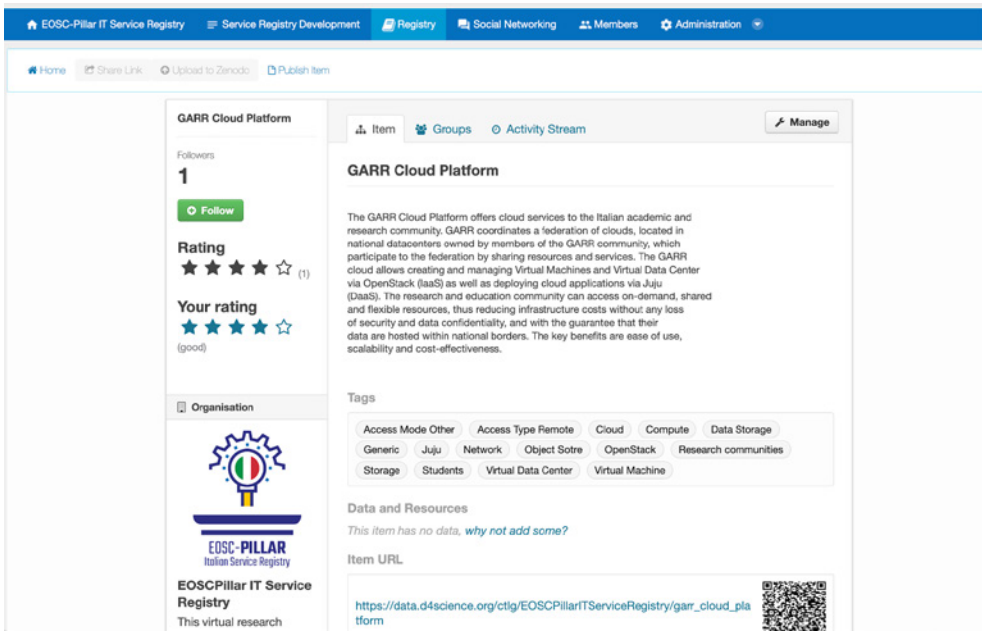
Recently, this document has been presented to other EC-funded projects related to EOSC during the EOSC-Pillar workshop "Incorporating National and Thematic Service and Resource Catalogues into the EOSC", held on July 5th 2021.

Feedback from external projects and initiatives is expected to improve the document.

Some useful pictures which may be included into the booklet are aside:



EOSC-Pillar
Service Registry Model



EOSC-Pillar Italian
Service Registry Prototype

12 https://repository.eosc-pillar.eu/index.php/s/o8H6LGbGDd3fpZF

EOSC Training

Emma Lazzeri CNR & Inge Van Nieuwerburgh UGent

EOSC-Pillar is immersed in an environment of existing projects and initiatives that represents a unique opportunity for sharing, engaging and efficiently collaborating towards the common goal of building the European Open Science Cloud federated infrastructure for FAIR data. In this context, training and support are essential activities to facilitate Open Science practices adoption and promote FAIR principles toward the various stakeholders. Like the other INFRAEOSC-05 projects, EOSC-Pillar has a strong training component oriented toward the national stakeholders.

To support these activities EOSC-Pillar can build upon the results achieved by other initiatives in Open Science training thanks to its network of Partners that are connected and involved in other initiatives and projects such as the Community of practice of training coordinators, RDA groups on training, GoTRAIN, OpenAIRE, FOSTER.

EOSC-Pillar therefore counts on existing and past training initiatives and projects to achieve its goal of setting up support and training activities facilitating the diffusion and adoption of mainstream standards and approaches for FAIR research data management, and an efficient uptake of the EOSC services in the region covered by the EOSC-Pillar partners.

During year 2 of the project, the Task 5.4 - training modules on FAIR-oriented research data management tools and solutions - continued its mission to deliver training events on EOSC and Open Science related topics to different stakeholders. From July 2020 to July 2021, the project organised 10 training sessions delivered to the national communities in Belgium, Austria and Italy, and trained more than 600 participants. Due to the COVID-19 pandemic, the training was delivered remotely. All materials produced were openly made available through Zenodo and the EOSC-Pillar training and support catalogue. Training activities targeted different research communities (COVID-19 researchers, Earth and environmental sciences and Humanities) and were designed in collaboration with the relevant ESFRI Research Infrastructures in the fields. Training events were also co-designed with research performing organisations and universities with the aim of training different roles inside the institutions, from research support staff to researchers at various career stages.

EOSC-Pillar T5.3 focused on Research Data Management (RDM) standards and on creating awareness in different data stewardship approaches in order to harmonize support offered to researchers. First the requirements for the EOSC-Pillar Research Data Management Training and Support Catalogue have been drafted and subsequently designed and implemented leveraging D4Science technology. The catalogue adopts specifications for metadata elements and values based on recent outputs from the WP6 (FAIR Competence Centre) of the FAIRsFAIR project. The catalogue is intended for research support staff (e.g.,- data stewards, project support, librarians, researchers). The work was subjected to a concentric feedback cycle going from the internal EOSC-Pillar community over to the Belgian community and then reaching out to a wider community. After the received feedback, a training workshop was held jointly with T5.4 and the catalogue was published as a first production release[13]. As of the first release, continuous update and improvement is organised. The activities were promoted during various events, which included among others Belgian Open Science webinars during the Open Access week 2020 and a co-organised FAIRsFAIR / EOSC-Pillar workshop in the spring of 2021. Furthermore, the documentation for promoting FAIR practices and the support to FAIR oriented data stewardship is reported in "D5.4:FAIR-oriented research data management Support, Training and Assessment Activity Report". Engagement with relevant initiatives is taken up, eg. through the OpenAIRE Training Coordinators Community of Practice

13 https://eosc-pillar.d4science.org/web/eoscpillartrainingandsupport

National Initiatives Survey

Lisa Hönegger, Anita Bodlos and Marie Czuray, UNIVIE

Focussing on Researchers and Building upon the "National Initiatives" Survey

The "National Initiatives" (NI) survey provides a snapshot of the state of national initiatives on open data and services in the five EOSC-Pillar countries. We gathered the survey data as a basis for other project activities, as well as a basis for evidence-based discussions and decision making in order to support the EOSC implementation. The "National Initiatives" survey focused on stakeholder groups on the implementer side, including research performing and funding organisations, as well as data and service providers – in order to get insights into the current research infrastructure landscape. This first phase dedicated to conducting and analysing the NI survey was completed by the end of the first year of the EOSC-Pillar project.

Analysing the Needs of Researchers

We were conducting a survey among members of user communities, their usage of currently available services and the maturity of these services. To analyse researchers' needs and practices, we conducted and will continue to conduct qualitative interviews focusing on data re-use by researchers.

In more detail, we developed a research design to assess the needs and difficulties researchers face when re-using data. Therefore, we gathered input and feedback from projects working on similar objectives (EOSCSecretariat.eu and FAIR Data Austria), especially during designing the semi- structured questionnaire. Following the FAIR principles, we asked what hinders and what facilitates the findability, accessibility, interoperability and reusability of data from a researcher's perspective. The aim is to gather insights on how re-using data and ultimately EOSC can be designed in a user-friendly way. To this end, we conducted qualitative semi-structured interviews with researchers of different disciplines, different career-stages and based in different EOSC-Pillar countries. Hence, the output will include views of a wide range of researchers.

In the future, the task plans to analyse the interviews using a method of qualitative content analysis. Based on the results, the task will revise the initial questionnaire and conduct new interviews and plans to continue these interviews or similar activities designed to gain insights from researchers as well as engage researchers in the remaining project lifetime.

Building Upon the "National Initiatives" Survey

EOSC-Pillar organised several webinars on Open Science policies and provided input from the NI survey results. The panel discussions by experts provided new views and arguments related to these topics. Another continuous activity will be the support of future EOSC projects and FAIR related activities with the use of the data gathered in the NI survey. This may include thematic specific activities (such as webinars, workshops, training, stakeholder engagement activities, etc.) related to the surveys, which will foster discussions and bring new findings and recommendations for the implementation of EOSC.

EOSC-Pillar Use Cases

EOSC-Pillar Work Package 6 Delivers Use Cases to Analyse Different Tools and Services for the "FAIRisation" of Data and Services and Type of Governance According to the Community Needs.

This work package collects use cases based on the requirements of scientific communities. Pilots run to validate the proposed solutions, which will be trans-national by design and general enough to be extended to other communities with minimal changes.

Download the report on the State of the Art and Community from Use Cases

The pilots include resources and cost study, in order to understand their feasibility and propose a viable business model for the resulting services. The selected use cases are connected to real scientific production and intent to improve the usage of data from a FAIR perspective.

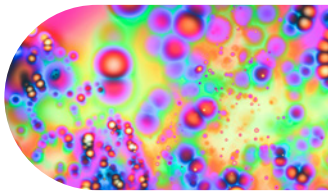
The need for "data FAIRisation" is important, diversified and specific for each community (nanotechnologies, environment -ocean, atmosphere, continental surfaces, health, humanities, biodiversity, solid earth). They are more or less structured and have developed tools and services adapted to their specific needs and constraints. The objective is to prepare scientific communities to be involved in EOSC.

The initial set of scientific use cases will be expanded through an open call for participation, inviting communities and their developers to bring forward thematic services to enhance the portfolio of the project and the EOSC Portal.

The main objectives of this work package are to:

- Analyse specificities of the scientific communities but also cross-cutting needs
- Integrate different data from all components and domains and facilitate access, integrated and intelligent treatments and diffusion
- Identify important tools and services from the use cases
- Identify gaps and needs that can be fulfilled by the future EOSC
- Expand the set of use cases through the launch of an open call for communities

Defining Procedures and Services to Enforce Data Provenance for Thematic Communities and Beyond



Due to data exploration complexity, provenance management is a key component in order to guarantee scientific data discovery, reproducibility and results interpretation. Provenance management should be able to define a set of metadata able to capture the derivation history of any stored data, including the original raw data sources, the intermediate data products and the steps that were applied to produce them.

To enable a well-defined data provenance in the scientific experiments workflow, from data production to data usage, this task will:

Elaborate cross-domain, FAIR-oriented procedures and recommendations to enforce data provenance by taking into account needs coming from various communities such as nanoscience/soft matter, material science and climate change in the field of environmental science;

Implement the defined procedures and recommendations defined by developing the proper adaptations/extensions on top of existing scientific data services, made available by the participating partners at national level.

- WEBSITE LINK
  ENES Climate Analytics Service
  TriDAS
- USE CASE
  https://bit.ly/3yE7eiv
- EMAIL
  cozzini@iom.cnr.it
  fabrizio.antonio@cmcc.it
  kulueke@dkrz.de
  kindermann@dkrz.de
- COMMUNITIES
  Climate science
  Material science/Nanoscienceì
- PARTNERS

Challenges Addressed

The initial challenge is to address and solve the provenance of data within the material science and climate science community.

For the material science community, we focused on the wide community behind the NFFA Europe Project where scientific data are collected from more than 150 experimental techniques. We planned to define general procedures to provide specific guidelines and recommendations on how to manage the important aspect of data provenance for NFFA Europe users community and beyond.

Climate research makes use of lots of data coming from the modeling and observational climate communities. In this domain, provenance management plays a key role both for numerical end-to-end simulations at the data center level as well as in the inner data analytics workflows.

Provenance enforcement procedures were identified, which will contribute to the climate data science notebooks of Use Case 2 (Agile FAIR data for environment and earth system communities). The provenance procedures can later be generalised for other domains. In addition, the approach will include a discussion of possible use cases of PID collections in this context and thus provides a link to Use Case 8 (Virtual definition of big datasets at seismological data centres according to RDA recommendations).

To account for in-depth provenance support within the climate processing services, a second-level provenance management complying with the W3C prov standard has been elaborated, thus addressing open science challenges (reproducibility, re-usability, etc.) at a finer granularity.

Benefits Through EOSC-Pillar

The material science use case will benefit through EOSC-Pillar by the fact that NEP user community and the project itself can establish a strong connection with EOSC network, favouring a continuous exchange among the two communities. Such exchange is of mutual interest: on one side (NEP) allowing the project and the data services built within the project to be EOSC- compliant. On the other hand, EOSC benefits from a large and committed user community that should provide useful suggestions and case studies in the overall EOSC implementation.

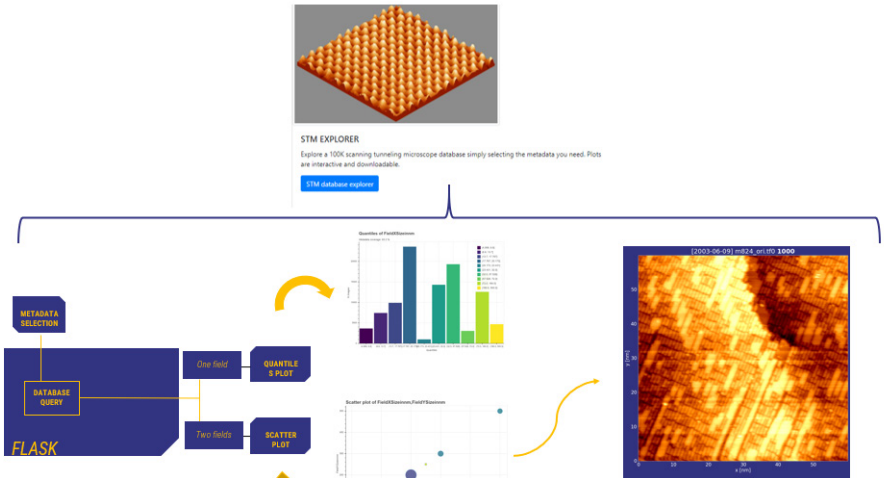
The climate use cases are built on top of the ENES Climate Analytics Service, which is the server-side compute infrastructure exploited in Use Case 2. ECAS, one of the EOSC-hub thematic services, allows performing data analysis experiments on large volumes of multidimensional data through a PID-enabled, server-side, and parallel approach. In this way scientific users can run their experiments and get fine-grained provenance information captured at the second level of a data analytics workflow based on W3C prov specifications. This allows retrieving the data lineage

of an object including the entire analytics workflow associated to it, which is particularly worth towards data discovery and the FAIR Reproducibility principle.

Highlights - Results as of May 2021

Material Science/Nanoscience Case Study

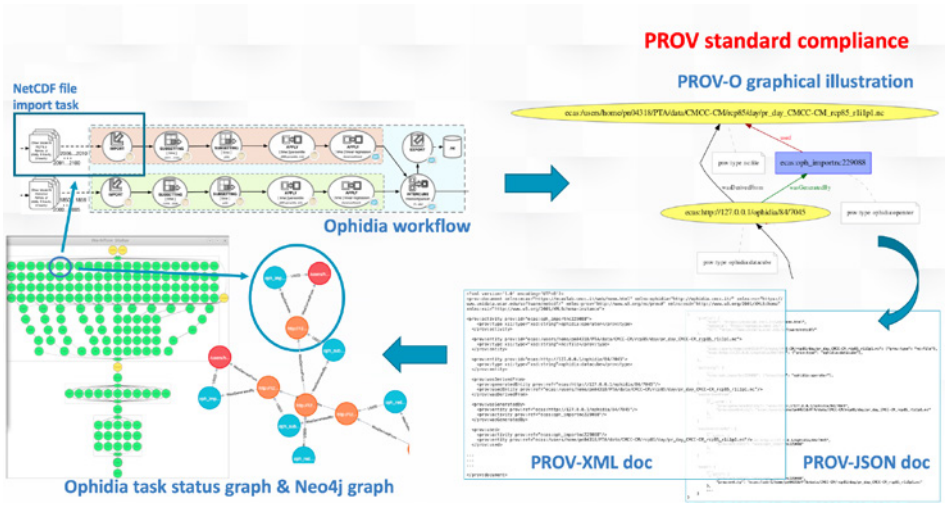
The case study within the community of material science/nanoscience concerns the creation of a database to store the scientific images metadata and the development of a web service to visually explore these metadata through interactive tools. The objective was to organize the original STM dataset (STM-Scanning Tunneling Microscopy), in a more structured and convenient dataset to allow researchers to build the provenance of data. Instrument metadata was extracted for every image of the dataset. The website was then built around the selected metadata fields, which are searchable and presented through graphs. Finally, the website was refined by linking the metadata to the underlying images so that researchers can visualize them directly on the browser without the need to use custom software. This tool, named STM Explorer, is integrated on the Trieste Advance Data Services website (TriDAS).



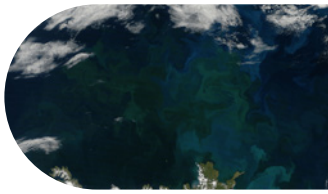
Climate Science

In the climate science domain, the multi-model Precipitation Trend Analysis (PTA) was selected as a pilot case. It has been implemented as an ECAS analytics workflow and executed on 11 climate models from the CMIP5 experiment for a total of about 200 tasks. Based on the information tracked by the Ophidia analytics engine about the executed tasks, scientific users running the analytics workflow can retrieve the corresponding provenance documents (complying with the W3C prov standard) in XML, JSON or graphical format through the Python application developed in the task as an extension to the existing ECAS modules in order to fully address FAIR-oriented provenance management.

An additional use case was identified in collaboration with Use Case 2. It specifically works on providing simple libraries to use in a JupyterHub environment, thus enabling scientists to generate W3C prov standard compliant provenance descriptions accompanying their jupyter notebook based analysis results. The approach to pre-define provenance templates[1] is evaluated with respect to simplifying this approach for scientists.



Agile FAIR Data for Environment and Earth System Communities



Earth Environment Sciences & Geosciences require a large panel and volume of data from satellite, in-situ observations and climate models, that are managed and preserved separately in domain-dependent repositories of national or European infrastructures. As the data sources are widely distributed, it induces real difficulties to achieve inter-processing and integrated uses of the data for comprehensive studies at domain or cross-domain level.

In this context, the goal is to offer services that facilitate and speed up access to the distributed data sources and provide a web-based processing environment for Data Science notebooks supporting the Pangeo community platform. This use case aims also to enhance data discovery and data access, relying on existing services provided by the Earth Sciences communities such as in France, the Research Infrastructure Data Terra, and in Europe, the consortium of Environment Research Infrastructures (ENVRI), the climate community, and beyond.

- EMAIL
  pillar@ifremer.fr
- USE CASE
  https://bit.ly/3iInVDA
- LEARN MORE
- COMMUNITIES
  Environmental science
  Geosciences
- PARTNERS

Challenges Addressed

The web-based processing environment proposed by the use case has to address the needs of two types of user of Data Science notebooks on Earth environment & Geosciences data :

- Scientists, who want to run ready-to-use data processing “Pangeo ecosystem-based” scripts on their temporal and geographical areas of interest;
- Data analysts, who want to develop, test and run their own data processing “Pangeo ecosystem-based” scripts adapted to their needs.

The first challenge to tackle is to design and implement the IT infrastructure (services and resources) underlying this virtual environment and that enables to speed up and facilitate access to data, taking into account the specificities of data sources, i. e. large volumes of data from distributed repositories.

The second one is to enable users to easily discover cross-domain data collections and services, even if these resources are distributed and based on different metadata and ontologies.

Benefits Through EOSC-Pillar

The use case is built on top of IT services and resources proposed by national infrastructures involved in the project:

- HPC and cloud computing resources from use case and Work Package 7 partners
- iRODS services from EOSC-Pillar partners, to ease and fasten data access from the web-based processing environment by synchronizing data from the distributed data repositories to the computing resources
- D4Science Virtual Research Environment to run Data Science Notebooks
- Software Heritage solution from the dedicated EOSC-Pillar use case, for the archive and sustainability of the source code of this use case’s Data Science notebooks
- Connection to Work Package 5 data layer services to demonstrate the cross-domain metadata catalogue enabling the data discovery. The first scenario will connect to D4Science and further scenarios foresee connecting to the Federated FAIR Data Space.

EOSC-Pillar also gives the opportunity to test cross-domain and transnational interoperability, as this use case gathers multi-domain data repositories from France, Germany and Italy.

Highlights - What Has Been Achieved?

The overarching goal is to demonstrate what EOSC could offer to the Earth environment & Geosciences communities, i.e. a cloud platform to run big data analysis in Europe as an alternative to using large private providers for storage and computing. To this end, the use case focuses on Pangeo as it is a community platform for big data (Geo)sciences oriented toward Python scripts developers, that is (1) fostering collaboration around the open-source Scientific Python ecosystem and (2) involving many relevant technologies: HPC, containers, notebooks, advanced data structures ("Data Cubes") for efficient access, remote access to data.

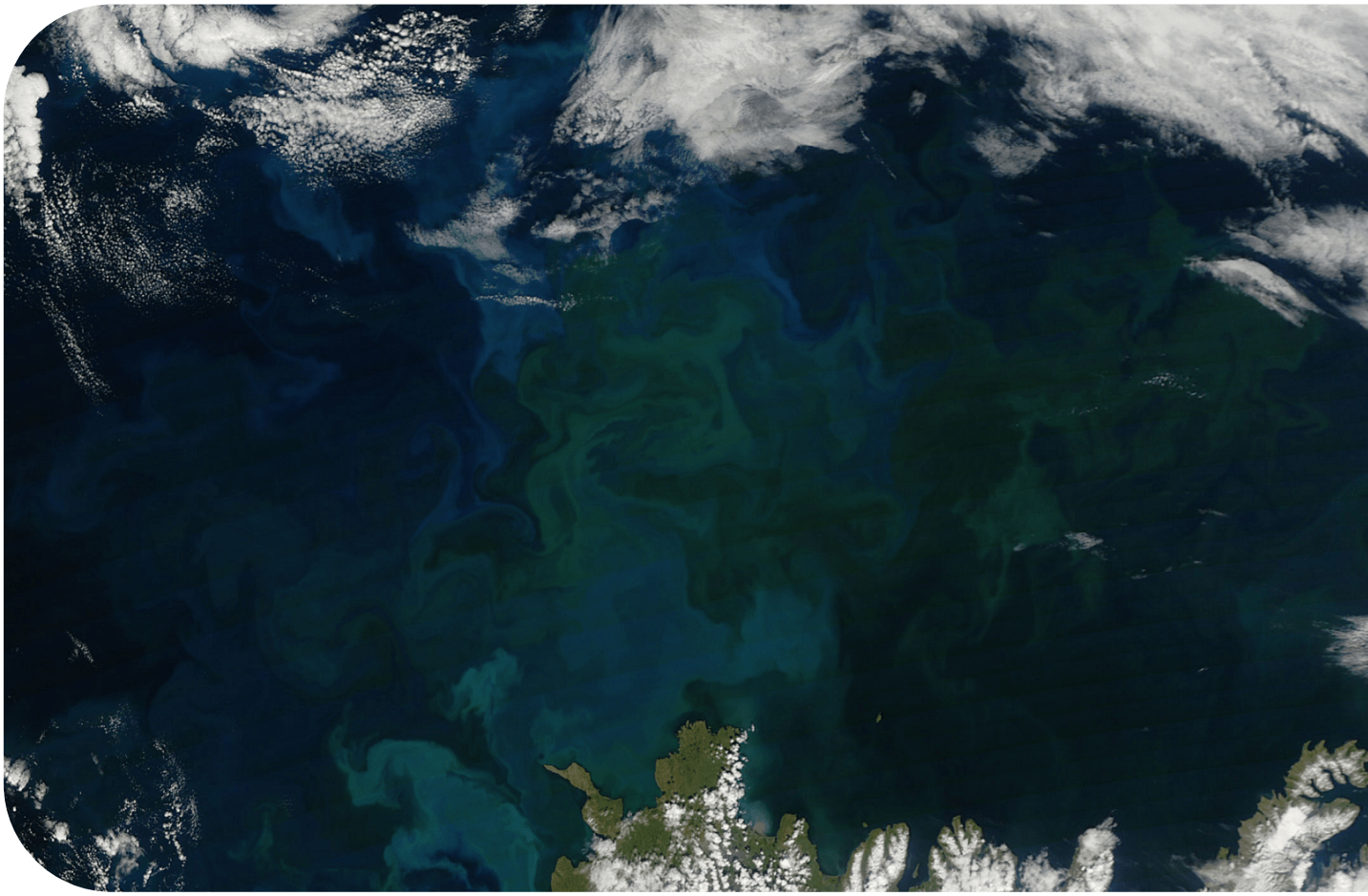
As the use case is ambitious to design and implement, the partners chose to split it into three sub-use cases, in order of priorities

- Data Science Notebooks, to offer a web-based processing environment for scientists and data analysts within the ecosystem of Pangeo
- Data services, to speed up and facilitate access to data of repositories from several domains,
- Discovery services, to provide a cross-domain catalogue

In addition, this use case aims to go as far as possible in the Proof of concept within the EOSC-Pillar project in collaboration with the technical Work Packages 7 and 5.

Demonstration Videos

- Data Terra: cross-analyse in situ (Argo) data and Copernicus Satellite data and interpolate sea level anomalies (Using Xarray, Dask and cartopy)
- DKRZ: Find, Analyze and Visualize Climate Data (using Intake-ESM, XArray and hvPlot)
- CMCC: Climate data processing and Visualization (using Xarray, Dask and cartopy/matplotlib)



Integration of Data Repositories Into EOSC Based on Community Approaches

The agriculture, food and environment research community faces many challenges common to all: Easily find and publish data, preserve them, and facilitate their treatment and analyse through computing solutions.

Examples of needs:

"I need to integrate innovative services that allow researchers to analyse data and publish it easily. Our information system must guarantee long-term storage of experimental data" - Engineer, Phenome-Emphasis community

"I need easy access to the data produced on the experimental platforms. I need to correlate various datasets. I need easy access to publish curated datasets or models." - Research Engineer, Phenome-Emphasis community

To address these needs, this use case aims to create a flexible federated research data ecosystem for the agrifood community through four aspects:

- long term data preservation,
- connecting data repositories,
- virtual research environments,
- cloud computing.

- EMAIL
  task3wp6eosc-pillar@groupes.renater.fr
- USE CASE
  https://bit.ly/3jYdpI6
- COMMUNITIES
  Environmental science
  Generic
  Life science
- PARTNERS

Challenges Addressed

- Data findability and reusability
- Data integration
- Data processing
- Reproducibility

Benefits Through EOSC-Pillar

By using the EOSC-Pillar Federated FAIR Data Space (F2DS), data providers and repositories will be able to make their data findable, accessible, and reusable by the whole community within the context of EOSC. As a direct consequence, this task will enable and/or increase interoperability among the repositories. Furthermore, the use case will leverage EOSC data services such as B2SAFE in order to implement long-term preservation of the institutional data repositories.

With EOSC distribution, the community as a whole will gain access to a research environment on which to process, analyse and visualise data in-situ with appropriate compute infrastructure, without the need to download them first, fostering collaborations and cross-fertilisation.

Highlights - What Has Been Achieved?

- Widening the agrifood initial scope through collaboration with partners.
- Deployment of a Virtual Research Environment.
- Deployment of OpenStack over INRAE and France Grilles infrastructures.
- Provisioning of Kubernetes Clusters based on INRAE and France Grilles infrastructures.
- Deployment of JupyterHub on INRAE infrastructures.
- Deployment of renku on INRAE infrastructures.
- Mapping between CINES archiving tool (VITAM) and DataINRAE metadata.
- Creation of Data INRAE openAPI definition to help integrate Dataverse based repositories data in the Federated FAIR Data Space (F2DS).
- Connected Galaxy to Dataverse.

Next Steps

- Setup a connector between CINES archiving tool (VITAM) and Data INRAE for automatically archiving Data INRAE's data in VITAM.
- Connect INRAE jupyterHub to the D4Science platform as a Jupyter notebook provider.
- Developing a connector between Dataverse and INRAE's Jupyterhub and renku instances.
- Developing a connector between Dataverse and D4science VRE.
- Setup interoperability between Fraunhofer Market Place and F2DS.
- Setup interoperability between Fraunhofer Market Place and Dataverse based repository.

Software Source Code Preservation, Reference and Access

Leveraging the experience of Software Heritage, this task aims at designing and pilot a solution for the preservation of massive collections of software source code (billions of source code files with links to publications) into EOSC eTDR service (European Trusted Digital Repository). More specifically, as part of this use case EOSC/Pillar will:

- Develop an API that allows platforms such as open access paper archives to archive in Software Heritage research software in source code form,
- Develop an API that allows anyone to retrieve and access archived source code artifacts,
- Standardize a schema of persistent identifiers (PIDs) that allows to reference billions of source code artifacts,
- Integrate the above APIs and Services with EOSC eTDR,
- Develop a pilot to fully replicate the software archive onto existing EOSC infrastructure.

- WEBSITE LINK
  Software Heritage
- EMAIL
  david.douard@softwareheritage.org
  zack@upsilon.cc
- USE CASE
  https://bit.ly/3yE7eiv
- COMMUNITIES
  Computer science
  Reproducibility
  Research software
- PARTNERS

Challenges Addressed

A large part of the technical and scientific knowledge that is being developed today resides in software. The preservation of this universal body of knowledge has become as essential as preserving research articles and data sets. Software preservation is a pillar of reproducibility, because software is used in essential ways during all phases of research in all fields of science. To be able to reproduce an experiment, knowing the exact version of the software used is essential.

Software Heritage will ensure availability and traceability of software, providing the missing vertex in the triangle of scientific preservation, together with open data and open access.

The first addressed challenge is to offer the research community a way to persistently and uniquely identify any piece of software source code. The second addressed challenge is to allow EOSC members to archive source code artifacts in the long term, thus helping reproducibility.

Long-term preservation guarantees will be achieved by replication, encompassing the main Software Heritage archive, its network of mirrors, and by depositing a copy of the Software Heritage Archive in the CINES long-term archiving solution (Vitam).

Benefits Through EOSC-Pillar

Thanks to EOSC-Pillar, users of the archival services offered by Software Heritage will benefit from integration with other common services and infrastructure offered by EOSC-Pillar.

One example of this is authentication and authorization (AAI). There will be no need to create dedicated accounts on the Software Heritage infrastructure to access the provided APIs: any identity provider integrated into EOSC-Pillar could be used.

Second, thanks to ESOC-Pillar Software Heritage got access to collaboration opportunities with partners interested in replicating the archive and their infrastructure, such as CINES with Vitam. In the future, other interested partners will be able to do the same, building on top the accrued experience.

For the future it will also become possible to partner with providers of computing resources, enabling research groups to conduct massive large-scale experiments on the source code artifacts archived by Software Heritage.

Highlights - What Has Been Achieved?

Results as of May 2021

- A persistent and intrinsic identifier schema (SWHID) for software source code artifacts has been specified and its adoption in the software research community is growing.
- The Deposit API, allowing partners to deposit their source code artifacts, is available and ready to use. This can be coupled with the AAI integration provided as a result of EOSC-Pillar Work Package 7 for easy integration of the deposit with services provided by EOSC partners.
- The Web and Vault APIs allows anyone to retrieve a source code artifact from the Software Heritage Archive starting from its SWHID.

Work in Progress

- Replication of the Software Heritage Archive into CINES' Vitam archiving service is under development. It will ensure that every source code artifacts present in the Software Heritage Archive, including those deposited by EOSC partners, will be periodically archived with high reliability and long term availability.

FAIR Principles in Data Life-Cycles for Humanities

This task aims to identify and develop use cases based on Social Sciences and Humanities (SSH) communities engagement. In order to do that, we will rely on consortia funded by Huma-Num and other partners from DARIAH (e.g. Italy and Germany). Another focus will be done on the link between data and publication. HAL, the French national open archive created by CCSD, provides a specific portal for SSH communities to deposit and deliver their publications in open access.

CCSD will work with Huma-Num to link publications on HAL to research data in data repositories, especially Nakala, the data repository from Huma-Num dedicated to SSH. This case will be a model for linking with other data repositories used in SSH communities. Finally, the idea is to liaise with CO-OPERAS GO-FAIR Implementation Network to ensure that SSH needs are correctly taken in account in order to facilitate the integration of SSH in EOSC.

- EMAIL
  adeline.joffres@huma-num.fr
  benedicte.kuntziger@ccsd.cnrs.fr
- USE CASE
  https://bit.ly/3iPBAcc
- COMMUNITIES
  Humanities
  Social sciences
- PARTNERS

Challenges Addressed

SSH are very fragmented and diverse. Presenting use cases representatives is a big challenge per se. So, gathering representatives of use cases from the CO-OPERAS community (SSH, scientific communication) and SSH in France is part of this work (Cf. MS24)

Our main focus represents in itself a challenge and a need widely shared in SSH communities: linking scientific publications with (raw) data stored in a structured and "FAIR" way in a secure repository. We thus decided to work on a POC consisting in linking publications in HaL (French Open archive) to data stored in the Nakala French repository dedicated to SSH (MS25). The implementation of this POC will address various core questions to current research in SSH, as you can see in the presenting video (embed?).

Finally, another big challenge is to imagine next uses and evolutions of the POC which could be replicated in other domains.

Benefits Through EOSC-Pillar

We can identify two main topics that SSH communities will benefit from the EOSC-Pillar project:

- Engaging a deep collaboration between two national infrastructures for SSH that will surely have a following and will allow a better comprehension and integration of common interests and needs
- Contributing to promote and make accessible Open Science and FAIR data processes and services
- Implementing new services to be deepened later for French SSH communities

Highlights - What's Been Achieved?

Next steps

Next challenge in parallel with the central Use Case 5 is to study in more detail the links and possible collaborations between the use case on Lidar data and WP7.

If the main highlight remains the implementation of the POC between the HAL and NAKALA repositories, various extensions are envisaged:

- The first step is to build the relationship between the publications deposited in HAL and the data deposited in Nakala, using the APIs available in each of the repositories. The relations thus created will be displayed, exported and harvestable.
- A second step will consist, if possible, in allowing the simultaneous deposit of publications and data in the same repository, HAL, and then transferring the data into Nakala, creating the relationship between the two repositories automatically, on the model of what has already been developed in HAL for software codes, in partnership with the Software Heritage repository. It is also conceivable, by means of a survey of users of the two data repositories, to deepen the bi-directional character of the "simple" link initially created, as well as its visualisation.

Exploring Reference Data Through Existing Computing Services for the Bioinformatics Community

Galaxy, one of the best-known workflow management systems for bioinformatics, aims to make computational biology accessible to research scientists that do not have computer programming or systems administration experience.

How can scientists connect this powerful tool seamlessly with many data sources? How can they do so in a coherent way using different instances of Galaxy? Can they run it locally or on a secured infrastructure that handles patient data? Can they compare the results of those different scenarios? Those are the main questions this use-case wants to address.

Building on top of existing French and Italian national services, the use-case will produce guidelines and best practices and implement a service prototype based on different scientific scenarios in order to:

- Allow access to reference data from different Galaxy deployments within the EOSC
- Facilitate the deployment of Galaxy instances close to the data
- Provide coherency between different existing Galaxy deployments
- Ensure health data security requirements are met throughout the process

- EMAIL
  gilles.mathieu@inserm.fr
  yosra.sanaa@inserm.fr
- USE CASE
  https://bit.ly/3m82ga2
- COMMUNITIES
  Health science
  Life science
- PARTNERS

Challenges Addressed

Galaxy is a widely used tool and comes in many flavours. One of the first challenges to address is the reproducibility and coherency of the different deployments to ensure that data analysis workflows produce the same results whatever the instance used. This means technical work within Galaxy itself, but also a global reflection on how to connect different data sources to Galaxy in a simple and coherent way.

Another challenge is the need to conform to data protection regulations concerning health personal data, by deploying Galaxy in a private, secured environment while still ensuring the data analysis workflow remains similar to its public counterpart.

Finally, we must find a way of providing access to the service to all users within the EOSC community through roles management and by integrating it into a global authentication framework.

Benefits Through EOSC-Pillar

This use-case relies on – and demonstrates the benefits of – EOSC-Pillar services at two levels:

- Discover and get data from an EOSC-wide federated dataspace. Through 4 different usage scenarios, the use-case will build on top of the FAIR Federated Dataspace (F2DS) to either collect metadata and information about datasets localisation or connect transparently to the source repositories through APIs.

Built on top of first-class services and resources provided through the project. These are mainly:

- Laniakea – Galaxy as a service provided by INFN. Laniakea[14] is a software framework that facilitates the provisioning of on-demand Galaxy instances as a cloud service over e-infrastructures.
- Data repository built by Inserm with the help and support of INRAe and CINES, around the Dataverse solution
- Cloud Galaxy instances provided by CNRS-IFB.
- Cloud computing resources provided by different partners
- The INDIGO-IAM authentication service provided by INFN

14 https://laniakea-elixir-it.github.io

Highlights

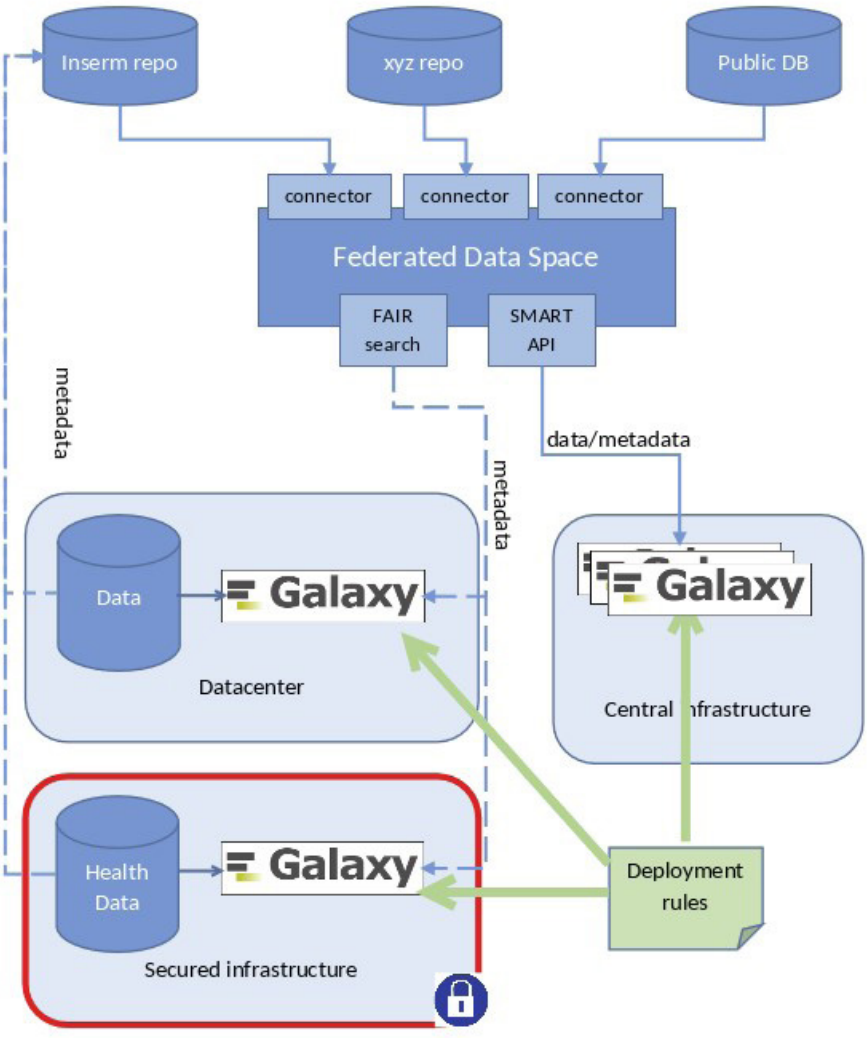
Results as of February 2021

- Clarification of requirements for the construction of the Federated Dataspace
- Gap analysis regarding “ready-to-use” services
- Global search on the type of targeted data
- Work started on analysing legal requirements
- Definition of 4 scientific scenarios
  - Public Galaxy Instance(s)
  - Close-to-the-data deployment
  - Health Data with restricted access
  - Reproducibility and verification
- Matching of the 4 scenarios with real-life examples, taken from the hCNV community
- First set of documents produced (Galaxy deployment state of the art and common best practices

Next Steps

- Connect source repositories, Galaxy and identified public databases to the Federated Dataspace (F2DS)
- Test, validate and implement the 4 scenarios
- Finalise reporting and documentation
- Test and validate provided services and solutions

Our 4 scenarios at a glance:



Suitable Data Formats for Seismological Big Data Provisioning Via Web Services

Seismology is one of the most advanced and well organised disciplines considering FAIR principles. With an International Federation of Digital Seismograph Networks (FDSN) providing guidelines and formal specifications for data formats and provisioning services, in order to foster the standardisation of data search, selection and retrieval, the community was able to develop an interoperable ecosystem of software clients and users.

The last few decades have been characterised by a large amount of new additional permanent and temporary conventional seismic stations becoming available. It is also already clear that there are some game-changing technologies under development. The demand for new dense observations using new technologies is advancing fast. This includes Large N deployments consisting of huge numbers of easy-to-install geophones, and fibre-optic based technologies (DAS), that are each starting to show their great potential in providing quality data with a wide spectrum of applications ranging from Tsunami early warning to Infrastructure monitoring.

Fibre-optic can be employed as to measure ground motion using the so called distributed acoustic sensing technique (DAS) with existing underground fibre-optic cables, present everywhere in cities and even in the seafloor (see Figure 1 below), up to tens of kilometres long, or even deploying dedicated short fibre-optic cables. Alternative usage of fibre-optic communication cables can be also placed along their repeaters (typically 50-100 km intervals) conventional sensors and only use the fibre-optic for communication, for example at the ocean floor using existing submarine cables. These techniques allow us to image the internal structure of faults with an unprecedented resolution, to infer creeping processes of faults at sub-micrometre steps, or to provide a global network of real-time data for ocean climate and sea level monitoring.

- EMAIL
  javier@gfz-potsdam.de

- USE CASE
  https://bit.ly/3m8c4Ro
- COMMUNITIES
  Geosciences
- PARTNERS

GFZ Helmholtz Centre Potsdam

KIT Karlsruher Institut für Technologie

Challenges Addressed

Although data quality and resolution of the techniques mentioned above are different, they have in common the potential to produce large volumes of data in a very short period of time due to both the extremely dense spatial and temporal resolutions. The datasets generated by these new technologies (e.g. DAS) would make it impossible to still use the same standard data formats and standard specifications for data provisioning services.

Seismological web services have been designed several years ago with particular types of user and data in mind, which are not exclusively what we see today, and these new acquisition techniques are a challenge for data centres and users.

The main aim in this use case is to keep FAIRness in the community by developing automated tools to standardise these datasets, and new implementation of the standard services capable of working with these new data.

Benefits Through EOSC-Pillar

Thanks to EOSC-Pillar, as well as developments in previous EOSC projects, users of dastools will benefit from integration with other common services and infrastructure offered by EOSC-Pillar. For instance, the Authentication and Authorization Infrastructure (AAI) integrated in EIDA/ORFEUS and provided by the B2ACCESS service hosted by Forschungszentrum Jülich.

Another benefit provided by the collaboration with Pillar partners is the capability to store the data requested in different cloud services (e.g. Nextcloud), or HPC facilities by means of services like Globus.

Fig. 1: A fiber-optic cable in the seafloor; after Jousset, P., 2019, Science

Highlights - What Has been Achieved?

Results as of May 2021

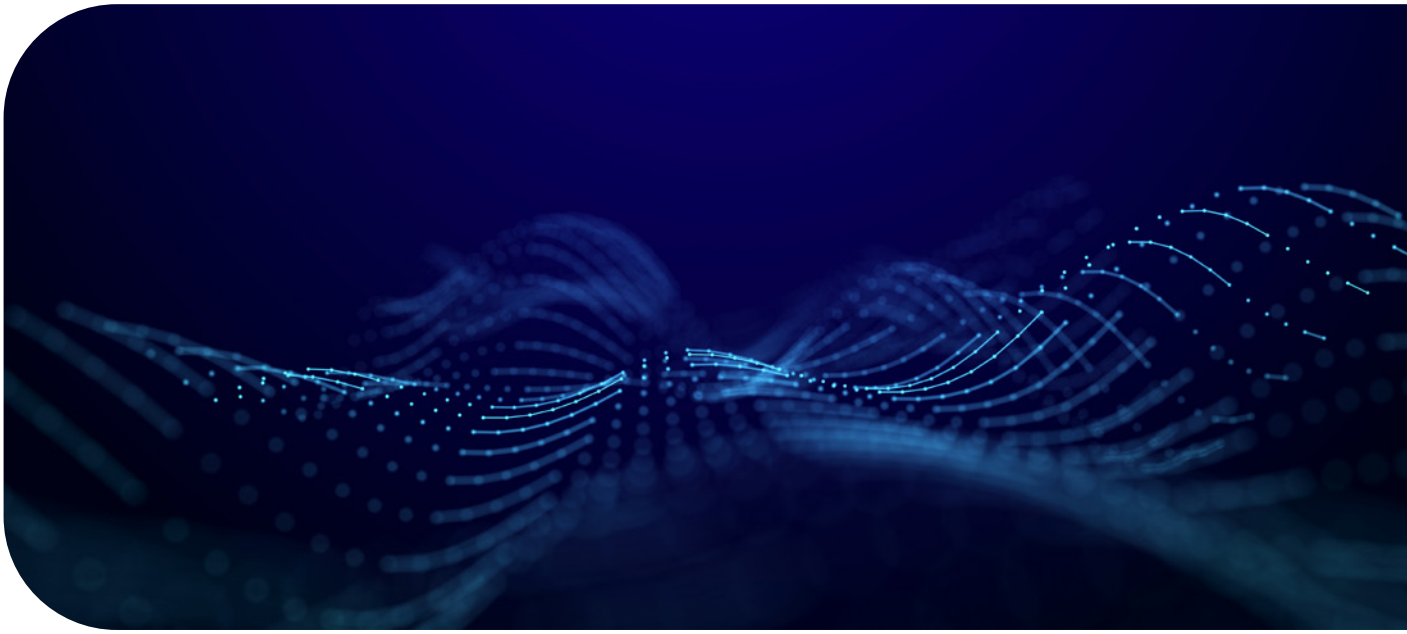
- Development and publication of dastools, a software package designed to standardise datasets from DAS experiments.
  - Capability to convert to standard formats.
  - Ready-to-use standard data provisioning web services on top of non-standard data format. Light weight and providing data in standard formats (e.g. miniSEED) and through standard protocols (FDSN-Dataselect).
  - Provision of ready to use containers to use the software.
- After almost two years of a joint effort with partners from two of the most important seismological data centers at international level (IRIS-USA, RESIF-France), we were able to provide a complete landscape analysis covering all aspects of the data life cycle for these novel datasets (see Publications below).
- We include in our analysis all aspects of the FAIR principles: not only the technical, but also the ones related to data management and even bureaucratic aspects of the international federation (FDSN) regulating the seismological standards for formats and services.
- Development of guidelines for a proper data acquisition and data ingestion into seismological archives. GEOFON is taken as an example, but the procedures remain as generic as possible.
- Special session at the General Assembly of the European Seismological Commission 2021 (ESC) in September 2021 to present and discuss results on the topics covered in this Task.

Future Steps

- On-going analysis of metadata changes needed in order to keep these new experiments within the standards, in cooperation with international partners. The final aim is to propose new standards where needed.

Publications

- Quinteros, J. (2021): dastools - Tools to work with data generated by DAS systems. V. 0.5. GFZ Data Services. https://doi.org/10.5880/GFZ.2.4.2021.001
- Quinteros, J., Carter, J. A., Schaeffer, J., Trabant, C., & Pedersen, H. A. (2021). Exploring Approaches for Large Data in Seismology: User and Data Repository Perspectives. Seismological Research Letters, 92(3), 1531-1540. doi: 10.1785/0220200390
- Call for abstract submission to the Session in the European Seismological Conference (ESC) in September 2021 https://www.erasmus.gr/UsersFiles/micrositе1193/Documents/SESSIONS_ABSTRACTS3.pdf



Virtual Definition of Data Sets According to RDA Recommendations

The management of digital objects remains an area of interest that crosses disciplines, institutions and infrastructures. In this context, the need for building aggregations or collections of such objects has become an essential element. Research data management practice requires not only to describe collections, but to make them actionable by automated processes to be able to cope with ever increasing amounts and volumes of data.

- EMAIL
  javier@gfz-potsdam.de
- USE CASE
  https://bit.ly/3k3S4g6
- COMMUNITIES
  Geosciences
- PARTNERS

GFZ Helmholtz Centre Potsdam

KIT Karlsruher Institut für Technologie

DKRZ Deutsches Klimarechenzentrum

Challenges Addressed

For many data centres, curators and dataset creators, it is difficult to assemble collections if there is the need for instantiating them, by replicating all their contents in one place. Sometimes this happens because of storage limitations, or because some parts of the dataset appear in many collections.

But the main limitation with existing solutions for the management of data collections is that they focus on describing collections and their semantics with metadata, but do not offer a full set of generic, machine-actionable CRUD operations on them.

A Working Group from RDA prepared specifications for a Research Data Collection system to overcome these issues.

This use case aims at making the Data Collections System, designed within the activities of RDA, generic enough and ready to be adopted by different communities within EOSC.

The system aims to:

- Provide CRUD operations (create, read, update, delete) for aggregations or data collections of different data objects as an essential element in the research data management practice,
- Describe the research data collections in a standardised way to make them actionable by automated processes in order to be able to cope with the increasing amounts and volumes of data.

Benefits Through EOSC-Pillar

This use-case relies on – and demonstrates the benefits of – EOSC-Pillar services as follows:

- Build a generic, multi-disciplinary Data Collections System based on the experience and requirements of 13 communities within the RDA Working Group,
- Improve the current system by revisiting its specifications based on the feedback we received from our partners in EOSC-Pillar,
- Allow the interoperation with the Federated FAIR Data Space and other communities involved in the Use Cases,
- Make the system available via containers to communities who would like to run them on their own.

Highlights

Achievements as of May 2021

- We revised the requirements from the seismological community for such a system.
- An implementation following this specification had been in use for some time at GEOFON, where more than 6000 collections and 1.5 million members for datasets had been pre-defined by the data centre operators. However, this system was only of internal use.
- After revising the requirements, we modified the system in order to make it completely generic and ready to be tested by other communities.
- One of the best aspects, regarding the resources needed to put the service in production, is that the members of a collection can be identified by PIDs (DOIs, ePIC), which means that almost no space is needed to define them. Within the context of this project, we added the capability to identify resources by URL, making the system independent from a Handle server in case that some resources (or the collection itself) needs to be identified.
- The identification of improvements to the current system by revisiting the RDA specifications is already advanced.
- A first set of requirements were collected from communities and other Use cases of the project. In particular, from Climatology, taking DKRZ as an example, and the Federated FAIR Data Space (F2DS).

Next Steps

- Deploy Data Collection System as a service to open it to more communities,
- Contact other stakeholders to foster usage of the system,
- Implement extensions and improvements to the RDA Recommendations. For instance, the automatic export to the Federated FAIR Data Space developed,
- Put the service in production.

Integrating Heterogeneous Data on Cultural Heritage

Heritage Sciences, i.e. the application of scientific experimental methods to the analysis of cultural heritage artefacts, produces a large quantity of numeric data that are only loosely related to the cultural object to which the analyses were applied. The lack of standard data models for the different technologies employed makes interoperability between datasets almost impossible. On the other hand, the same cultural objects and activities on them (studies, interventions, etc.) are documented in textual documents usually with very basic metadata.

This situation requires the intervention of a human to link the documentation of scientific analyses to the documentation of the cultural object, e.g. chemical analyses and physical to a study by an art historian; this in the end prevents data re-use and data-driven research.

- EMAIL
  franco.niccolucci@pin.unifi.it

- USE CASE
  https://bit.ly/3mo0n9F
- COMMUNITIES
  Cultural heritage
  Humanities
- PARTNERS

Challenges Addressed

The creation of a Linked Data system, as outlined above, requires the creation of a common semantic model covering both the scientific data and the content of text reports. This has been created as an application profile of CIDOC CRM, the standard ontology used for cultural heritage documentation, based on previous extensions such as CRM-PE, developed within the PARTHENOS project, and CRM-SCI.

Based on this newly-created schema, the data encoding may be obtained by uploading the numeric results of the scientific experiments together with their metadata to create the scientific side; and then uploading the text data annotated with the same ontology. The system will then create the links between the two components of the documentation system. However, the manual annotation of texts is cumbersome and time consuming, so a text mining tool based on machine learning to enrich the metadata of the texts which may be linked to the metadata accompanying the digital outcomes of the scientific analyses.

Benefits Through EOSC-Pillar

Sharing the tool and supporting the integration of scientific data with heritage reports in a cloud environment may foster cross-fertilisation between heritage professionals and scientists with data scientists. It may be a significant step forward in the digital transformation of cultural heritage and - at least - push forward the digitisation of heritage assets and their documentation, by showcasing the new opportunities opened by such integration.

Highlights - What's Been Achieved?

The system is fully working and has been tested on a set of archaeological reports/scientific data. To train the NER tool, reports were manually encoded by experts, and then the automatic annotation was tested on others, working pretty well. A merger set of training texts would be necessary to extend the usability to their fields or languages.

www.eosc-pillar.eu

@EoscPillar

/company/eosc-pillar/

bit.ly/3evkLxI

zenodo.org/communities/eosc-pillar/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 857650.