CNRS institutes and the EOSC – status and vision

Since the start of the EOSC related activities in 2016, CNRS has been a major partner in making EOSC vision a reality. In this context, the different implication of the institutes of CNRS in this process reflects the general approach: the EOSC has to be understood as a process, which aims to help and support scientific communities in their specific efforts. This should help to tackle the wide range of challenges we are facing in each scientific domain, from the FAIRisation of data to the improvement and opening of data and computing services through the seamless access and use of data and e-infrastructures. Therefore a general vision of CNRS vis-à-vis the EOSC must take into account the institutes' specific implication, research-driven strategy and requirements. We therefore organized individual discussions with the representative of each institute in the CNRS-EOSC working group. A short summary of the results can be found below, and overview of implication and needs is given in Table 1 and 2 at the end of this document.

It turns out, that CNRS is already involved in a large number of EOSC related projects: EOSCpilot, EOSC-hub, EOSC-Pillar, XDC, TRIPLE, ESCAPE, ENVRI-FAIR, SSHOC, AENEAS. In addition, there are projects and initiatives, which are closely related to the EOSC and where CNRS plays a major role, for example ELIXIR, PHIDIAS, GO-FAIR, OpenAIRE, EGI, and colleagues from CNRS represent France in various EOSC Working Groups (FAIR data, Architecture, and Landscape).

Although the expectations from colleagues in the different institutes vary, there are four main areas of interest emerging, in which our scientific communities hope to advance significantly thanks to the EOSC:

- **FAIRisation of data**, to enable data sharing and data reusage
- **Creation of trusted repositories** to provide platforms where FAIR data can be stored, managed, cured, and retrieved
- Uptake and consolidation to allow the use of already existing **data and computing services** in a larger context and with a long term perspective
- Seamless access to (heterogeneous) centralised and edge data and **computing infrastructures** across domains and technologies

Clearly, these goals are not only expressed in the context of the EOSC, but are also part of national and CNRS initiatives, as for example the FAIRisation aspect is addressed in the *Feuille de route du CNRS pour la science ouverte* and the GoFair French contribution. But the EOSC allows addressing a larger range of trans-national aspects in the European and international context. Clearly expressing CNRS research-driven needs and expectations is going to ensure that these key aspects are included and supported through the EOSC strategy on national and European level.

1. IN2P3

IN2P3 has been active in supporting the EOSC vision and its projects from the very beginning.
The particle and astroparticle community in IN2P3 has a long-standing expertise in FAIRisation of data, sharing of data on an international scale, and providing distributed computing facilities across borders and domains.
The main challenges here are now the virtualization of processing in order to use heterogeneous e-infrastructures, as well as the further development of grid and cloud solutions for BigData challenges of the next generation of experiments in the domain.
There is also a number of services which are used in parts of the community, which would benefit from being integrated into the EOSC portal.
At the same time, the nuclear physics community is adapting the FAIR principle for their data and the EOSC can help in this process.

1.1 EOSC project participation
EOSC-pilot (WP lead), EOSC-hub, EOSC-Pillar (WP lead), XDC, ESCAPE (project lead), EGI (council and executive board), EOSC Working Groups (Landscape and Infrastructure)

1.2 EOSC interests

- grid and cloud infrastructures in EOSC
- heterogeneous e-infrastructures in EOSC
- uptake of services into the EOSC portal

2. INC

Discussion : 5.11.2019, Marc Baaden, Volker Beckmann

Experiments within INC are often organised by small groups in a single lab. Often, there is no knowledge of the FAIR data concept at the researcher level. First need is to improve the FAIRness of INC data. Connected to that, there is a lack of common data standards and rules for metadata.
In addition, trusted repositories are an issue. Although in some domains they exist and are used (e.g. the PDB Protein Data Bank ; https://www.rcsb.org), this is not given in many fields of chemistry and a link between the existing repositories can not easily established.
In some cases, zenodo is used as a repository solution, but also Google-services are often a work-around.
Concerning computing, the most significant demand is for virtual machines on HPC systems in order to perform simulations.

2.1 INC EOSC project participation

Right now, INC is not involved in any EOSC project

2.2 INC EOSC interests

- FAIRisation of data,
- creation of trusted repositories,
- interoperability and interfaces between existing repositories,
- seamless access to HPC systems for simulations

3. INEE

Discussion: Sylvain Lamare, 7.1.2020 via e-mail

The global aims of the biodiversity field are to understand the underlying mechanisms of nature, document and capture the state and dynamics of ecosystems, and build predictive models for the future. This understanding is based on access to and use of data, models and analysis tools, produced in ever greater quantities, and used by diverse communities tackling different aspects of biodiversity from observations, collections, sampling and experimental data.
Through two structures (UMS), CNRS INEE engaged with other partners a large action devoted to: 1) create a single platform aggregating all biodiversity data and species collections starting with the PNDB and related initiatives (Recolnat, INPN, GBIF France) in France ; 2) develop scalable FAIR workflows and services to aggregate, analyse and increase the level of FAIRness of the data ; 3) increase the richness of metadata descriptions ; 4) develop competences in semantic modelling, promote best practices and emerging standards within the diverse institutes and universities across the francophone community who contribute data and resources to this shared resource, and 5) produce French language training materials, webinars, and an e-learning platform to increase the awareness of FAIR and the Internet of FAIR Data and Services.

3.1 INEE EOSC project participation
EOSC-Pillar, GO-FAIR, EOSC Working Group (FAIR data)

3.2 INEE EOSC interests

- FAIRisation of data
- DMPs
- Training/ involvement of researchers
- Trusted repositiories
- Shared services

4. INP

Discussion: through e-mail, Laurent Lellouch

INP is mainly engaged in the EOSC through its responsibility for the Photon and Neutron (PaN) facilities, ESRF, Soleil and ILL. In this context, FAIRisation of data is an important aspect for INP right now. A significant amount of work has been already invested on putting data management policies (DMP) in place, on assigning DOIs to data from the physics domain, and on metadata standards and catalogues, as well as on virtualized working environments for analysis (e.g. Jupyter notebooks). These services and the data

generated in these facilities are being made available to the consortium of users of PaN facilities in Europe, and it is planned to make them available to the broader scientific community through EOSC. INP is also involved, with INC, ILL, Soleil and a number of chemistry IRs, in finding ways to extend ILL' s DOI attribution system to other RIs, in devising discipline-specific and modular DMPs, as well as in discipline-specific and instrumentation-agnostic sample metadata describing the physical samples used in experiments. Connected to this is a requirement for data repositories, which extends to the wider INP community.
On the e-infrastructure side, the INP community has a strong interest in heterogeneous computing infrastructures to provide access to HTC and HPC facilities at the same time.

4.1 INP EOSC project participation
Right now, INP is not directly listed as partner in any EOSC project, but is involved indirectly through its TGIRs ESRF, Soleil and ILL, e.g. in PaNOSC and ExPaNDS.

4.2 INP EOSC interests
- FAIRisation of data
- Trusted repositories
- Heterogeneous e-infrastructures in EOSC

5. INS2I

Discussion: 3.12.2019, with Ali Charara (DI), Michel Daydé (IRIT), Denis Girou (IDRIS), Denis Veynante (MICADO), Pierre-François Lavvallée (IDRIS), Volker Beckmann

INS2I has been involved in the EOSC since the beginning through the EOSCpilot project. A main point of discussion is the possible integration of HPC systems within the EOSC and how this would relate to EuroHPC and PRACE.
In terms of FAIR data, colleagues at INS2I are rather on the users-side of data than on the production side. Common data platform needs are served by e.g. PerSCiDO at GRICAD, OSIRIM at IRIT and GAMATICA at LIMOS. For computing testbed resources, INS2I relies on SILECS, including FIT and Grid'5000
A main interest is the safe storage and sharing of algorithms and software. Here, the main platform used is Software Heritage, provided by INRIA. There is also some general interest in the management and the deployment of services in large-scale heterogenous distributed environments.

5.1 INS2I EOSC project participation
EOSCpilot, EOSC-Pillar

5.2 INS2I EOSC interests
- service sharing
- software repositories
- heterogeneous infrastructures

6. INSB

Discussion: 4.12.2019, with Daniel Boujard, Volker Beckmann

The themes of the EOSC, like FAIR data sharing and shared services, are vital for INSB's core topics. With two main domains, the study of living structures (like cells) and those of molecular biology now having a significant interest to work with each other, to share data, expertise, and services, biology is one of the fields most interested in an EOSC-type evolution.
The main link to the EOSC is the ELIXIR H2020 project, although ELIXIR itself not a EOSC project. The French partners in ELIXIR are coordinated through the IFB (Institut Français de Bioinformatique), which is funded by CNRS, INRA, INRIA, CEA, INSERM, universities, CIRAD, and the Curie and Pasteur institutes. Through the IFB, CNRS / INSB also participates in EOSC-Life, although CNRS is not a direct partner of this project.
The main demands in the INSB domain are on FAIRisation of data and of providing repositories at IFB to share data within biology but also with e.g. the medical sector and environmental sciences (on topics such as biodiversity). Following this, common services, provided through IFB and/or ELIXIR are the next step.
Concerning computing and storage infrastructure, the IFB partners provide these through the IFB portal, with a total of ~22,000 CPU and 11 PB disk space through federated clusters and clouds. On the infrastructure side, the next step is the connection of these resources in the EOSC through ELIXIR.

6.1 INSB EOSC project participation
EOSC-Pillar

6.2 INSB EOSC interests

- FAIRisation of data
- Trusted repositiories
- Shared services
- Coordination of ELIXIR / EOSC participation through IFB

7. INSHS

Discussion: 13.11.2019, Suzanne Dumouchel (Huma-Num), Volker Beckmann

INSHS has been involved in the EOSC and in the Open Science discussion from the beginning. Nevertheless, at the researcher level, there might be some knowledge about FAIR data principles, but the EOSC is rather not known. In terms of FAIRisation of data, projects have been conducted in the context of RDA, but FAIRisation remains the largest challenge.
Many of the EOSC related activities are channeled through Huma-Num and OpenEdition. There is a strong interest to integrate existing services of the French community into the EOSC portal, and these activities are e.g. part of the EOSC-Pillar, SSHOC and TRIPLE participation of INSHS. OpenEdition and Huma-Num are also coordinating OPERAS, the European Research Infrastructure for the development of open scholarly communication in the social sciences and humanities, which has also a strong link with the EOSC.

7.1 INSHS EOSC project participation
EOSC-Pillar, TRIPLE (coordinated by Huma-Num), SSHOC, GO-FAIR, OpenAIRE, EOSC Working Group (Architecture)

7.2 INSHS EOSC interests

- FAIRisation of data
- DMPs
- Trusted repositories
- Training/ involvement of researchers
- Service uptake into the EOSC portal, connection of catalogs of services on EOSC level (e.g. connection with Opidor)
- Controlled vocabularies

8. INSIS

Discussion: 8.11.2019, Fabien Godeferd, Volker Beckmann

The work at INSIS relies mainly on independent groups, which have no strong links concerning data sharing and data organization, although the scientific communities are well structured. The most relevant topic in this context right now is FAIRisation of data, starting with the need to have agreed standards and data formats in order to be able to then share data. Data management plans (DMPs) are used in some, but by far not all experiments at INSIS.
Certain platform initiatives are in place to provide trusted data repositories, e.g. in the context of fluid mechancis.
In terms of infrastructures, most relevant is the use of HPC resources through PRACE, and some usage of computing grid resources.

8.1 INSIS EOSC project participation
Right now, INSIS is not involved in any EOSC project

8.2 INSIS EOSC interests

- FAIRisation of data,
- DMPs
- Trusted repositories, storage
- Service uptake

9. INSMI

Discussion: 15.11.2019, Christophe Berthon, Volker Beckmann

In terms of FAIR data, colleagues at INSMI are rather users than producers of data. In that context, there is a strong interest in FAIR data, but not a specific need to help researchers at INSMI to make their data FAIR. Thus, there is some knowledge about the FAIR principle at the laboratory level, but little to no knowledge about the EOSC and its goals/opportunities.
There is, however, an interest in data repositories in order to store data that have been used in e.g. statistical analysis.

9.1 INSMI EOSC project participation
EOSC-Pillar

9.2 INSMI EOSC interests
- trusted repositories

10. INSU

Discussion: 13.11.2019, with Maryvonne Gerin, Jean-Pierre Vilotte, Volker Beckmann

INSU has been part of the EOSC initiative from the beginning and has a long-standing expertise within some of its pioneering research communities (e.g Astronomy and Astrophysics, Climate, Seismology) in FAIR data sharing, long-term data archiving, curation and open access. In this context, INSU can share their expertise with the EOSC community.
Building on the expertise of its pioneering communities, the main challenge INSU is facing today is to leverage FAIR data practices within and across the different disciplines and research practices in the institute, propagating, shaping and sustaining methods and services that have proved their value. This community-driven shaping and capacity-building strategy (e.g. the national research infrastructures Data Terra and CDS) requires coordination across national and international communities and research organisations. This is critical to address new and challenging inter-disciplinary and trans-disciplinary research practices, which stress the use and re-use of an increasing diversity and volume of multi-source and multi-type data (observation, simulation).
The interest at INSU in the EOSC is: long term multi-source and multi-type FAIR services and certified repositories; software platforms of services supported across edge and centralised data and computing infrastructures enabling data logistics across large-scale workflows (from high-end data streaming acquisition systems at the edge to centralised computing (HPC and Cloud) and data infrastructures; sustainable, trusted and long-term data and computing services on which the INSU community-driven shaping strategy can build on.
The main expectation here is for the EOSC to provide a framework for sustained software distributed platforms of services and standards. While standards can be mainly developed through the Research Data Alliance (RDA) initiative, the need for sustained services supported across centralised and edge infrastructures requires the existence of scalable and research-driven e-infrastructures within the EOSC.
Another expectation is for the EOSC to provide a framework which sets out the responsibilities of data creators, data users, research funders and publishers for data produced by publicly funded research, since abrupt loss of tools would be as deleterious as abrupt loss of access to data.

10.1 INSU EOSC project participation
EOSCpilot, EOSC-Pillar (WP lead), ESCAPE, ENVRI-FAIR, AENEAS, PHIDIAS, GO-FAIR, EOSC Working Group FAIR data

10.2 INSU EOSC interests:
- persistent services
- software platforms
- scalable e-infrastructure solutions

11. DIST

Discussion: through e-mail, Laurence El Khouri, Sylvie Rousset

CNRS has released its open science roadmap where the research data and the participation to European and International initiatives regarding open access to research data and publications, as well as data sharing and data management, are mentioned. The DIST is addressing subjects like data repositories and data management through working groups including CNRS Institutes. DIST is a stakeholder of the Ministry of higher education, research and innovation (MESRI) committees and working groups and thus participate to the definition of the French open science policies.
CNRS through DIST is an associate member of the OpenAIRE Legal Entity and thus addresses the necessary shift of scholarly communication towards more openness and transparency and implements innovative ways of to communicate and monitor research.
Several of CNRS DIST activities in support of the CNRS Open Science Roadmap and National Open Science Strategy deal with FAIRisation, and are thus strongly relevant to the EOSC. INIST is developing tools and services to facilitate a proper management of research data, with in particular among the OPIDOR services DMP OPIDOR, a tool for on-line creation of data management plans, and the attribution of Digital Object Identifiers, which are an essential element of the FAIR principles, through DataCite.
The development of data management/sharing culture and skills among all the stakeholders of the data life cycle is also a key building block of data FAIRisation. The DoRANum platform, developed by INIST and the Urfist Network GIS, enables self-training on data management and data sharing. The DIST is networking with the institutes through its STI correspondents.
The DIST is also co-leading the RDA France National Node, which has certification of data repositories as a priority, which is a key element for establishing trustworthiness. Workshops are organised to support repositories seeking certification. RDA France has a network of data correspondents in the institutes. More generally, it disseminates knowledge about RDA activities, many of which are very relevant to EOSC, in the CNRS (and French) community, and encourages the community to participate in the RDA. Specific actions can be engaged on topics of interest for the Institutes and EOSC-related projects for which RDA could bring input useful to EOSC.

11.1 DIST EOSC project participation
RDA Europe 4.0, Openaire, EOSC Working Group FAIR data

11.2 DIST EOSC interests:

- project preparation INFRAEOSC 07-2020 and related topics
- EOSC WGs, e.g. on Rules of Participation
- FAIRisation of data
- Trusted repositories
- Service uptake / consolidation
- Training

		EOSC projects									Related to EOSC					
		EOSC-pilot	EOSC-hub	EOSC-Pillar	XDC	TRIPLE	ESCAPE	ENVRI-FAIR	SSHOC	AENEAS	ELIXIR	PHIDIA	GO-FAIR	OpenAIRE	EGI	WG
	#French partners:	4	4	6	1	1	1	5	1	1	4	3	1	4	1	3
Institute	#participations															
IN2P3	6	X	X	X	X		X								X	2
INC	0															
INEE	2			X									X			1
INP	0															
INS2I	2	X		X												
INSB	2			X							X					
INSHS	5			X		X			X				X	X		1
INSIS	0															
INSMI	1			X												
INSU	7	X		X			X	X		X		X	X			1
DIST	1													X		1
CNRS	**15**	X	X	X	X	X	X	X	X	X	X	X	X	X	X	6

Table 1: participation of CNRS institutes in EOSC projects and in EOSC-related activities. WG = EOSC Working Group (Architecture, FAIR data and Landscape)

Institute	Trusted repositories	FAIRisation of data, DMPs	Service uptake/consolidation	(heterogeneous) e-infrastructures, HTC/HPC	Grid/cloud computing	Software repositories	training
IN2P3			X	X	X		
INC	X	X		X			
INEE	X	X	X				X
INP	X	X		X			
INS2I			X	X		X	
INSB	X	X	X				
INSHS	X	X	X				X
INSIS	X	X	X				
INSMI	X						
INSU	X	X	X	X	X	X	
DIST	X	X	X				X

Table 2: main interests in EOSC topics expressed per CNRS institute.