EOSC Future

D8.6

Procurement Plan for Commercial Data in Support of Research Data Spaces

The EOSC Future project is co-funded by the European Union Horizon Programme call INFRAEOSC-03-2020, Grant Agreement number 101017536

Version 1.0
October 2022

D8.6 / Procurement Plan for Commercial Data in Support of Research Data Spaces (discussion paper)

Lead by **GÉANT Association**
Authored by Dave Heyns (GÉANT), Hendrik Ike (Appendix A, GÉANT)
Reviewed by Ron Dekker (TGB) & Athanasia Spiliotopoulou (JNP)

Dissemination Level of the Document

Public

Abstract

This document is intended to be a working document by which various EOSC Future stakeholders can build a collective understanding of how commercial data (currently consumed by researchers) plays a role in the European research data spaces.

The document describes how Task 8.1 intends to demonstrate a mechanism by which commercial data required by European researchers can be distributed by the European Open Science Cloud. Commercial data is data resources that are produced/curated by commercial entities and are (currently) procured by research institutes in support of research activities. This may include background knowledge such as data sets, software, databases, etc. Specific examples of such resources need to be identified through a requirements and demand analysis, with initial focus on a single science discipline. Task 8.1 will work closely with the relevant science cluster and any other related EOSC projects to identify example data resources that are in relatively high demand by researchers in the selected discipline in order to demonstrate the sustainability of the distribution mechanism.

Once identified, the task will need to engage with the providers of the data resources through a market analysis, to understand the commercial models by which this can be accessed. Cost, terms and conditions and provider location will need to be understood so that a procurement construct can focus on compliance in terms of European Union Procurement Directives, and VAT regulations.

An elaborate procurement plan, possibly involving aggregating entities who can register the resultant pre-procured data sets as EOSC resources, will need to be carefully constructed based on the outcome of described analysis. Hopefully this plan can harness lessons learnt in the pre-procurement of other types of resources (digital services).

Registered data resources will need to be identified in terms of their thematic focus in order to be consumed by other discipline-oriented frameworks.

Version History

Version	Date	Authors/Contributors	Description
V0.1	19/10/2022	Dave Heyns (GÉANT)	Initiation – Proposed ToC – First draft
V0.2	23/10/2022	Dave Heyns (GÉANT), Hendrik Ike (GÉANT)	Final Draft
V0.3	24/10/2022	Dave Heyns (GÉANT), Hendrik Ike (GÉANT)	Updated version to incorporate comments
V1.0	26/10/2022	Dave Heyns (GÉANT), Hendrik Ike (GÉANT), Ron Dekker (TGB), Mike Chatzopoulos (ATHENA)	Final Version submitted to EC

Copyright Notice

Table of Contents

Table of Figures

Glossary

EOSC Future project Glossary is incorporated by reference: https://wiki.eoscfuture.eu/x/JQCK
In addition to the EOSC Future glossary, it is important to define a few specific terms referenced in this document:

Term	Definition
Aggregator	This describes an entity representing the aggregated demand (by the research community) for digital resources (both services and data). This entity would be responsible for the distribution of related resources via the registration of such on the EOSC marketplace. In terms of commercial resources (services/data), these would be procured on behalf of the aggregator who would then distribute them. Examples of potential aggregators would be the e-Infras/ERICs/NRENs.
Provider/Supplier	A commercial entity that produces/compiles/curates datasets and makes these consumable by researchers by means of a commercial model (procurement/ subscription). Examples of these could be pharmaceutical companies making clinical trial data available to the science community at a cost.

List of Abbreviations

Acronym	Definition
AI	Artificial Intelligence
API	Application Programming Interface
FAIR	Findable; Accessible; Interoperable; Reusable
CEDS	Common European Data Spaces
EOSC	European Open Science Cloud
LS	Life Sciences
PoC	Proof of Concept
RI	Research Infrastructure
RoP	Rules of Participation
SME	Small and Medium Enterprise
SPV	Single Purpose Voucher
SWD	Staff Working Document
NREN	National Research and Education Network

1 Executive Summary

Task 8.1 in EOSC Future has been mandated by the Commission to focus one-third of the project's commercial services adoption funding budget on demonstrating mechanisms by which commercial data used by researchers can be procured and distributed via the EOSC and the European data spaces.

Data produced, collected, and curated by commercial entities, including small and medium enterprises focused on the support of European research, is currently extensively used by the regional research community. At present, it is the burden of the researcher, or related research institute, to procure the licence to consume commercial datasets, possibly by means of a subscription.

Initially, the Task 8.1 team proposes to work closely with a single science discipline, through the related cluster project, to identify the most common commercial datasets currently used by their researchers. The procurement team on Task 8.1 will engage with the organisations distributing this data to better understand their commercial models, and the indicative cost (and terms and conditions) of a large community accessing their data resources. The procurement plan will be developed based on how the possibilities relate to the EU procurement Directives, and the physical location of the supplier and the research community in terms of VAT regulations.

The mechanism by which the data is pre-procured and distributed via entities able to register the resultant data resources on the various EOSC catalogues will need to demonstrate some level of sustainability if this mechanism is to add real value to the European Open Science Cloud and its various stakeholders and end users in future. The results will challenge current funding models, as there will need to be additional mechanisms allowing the re-procurement of these resources once the previous funding budget is exhausted. The rate of consumption of the licence supporting use the data resource will need to be carefully monitored in order to ensure that the availability of this resource is not interrupted, and re-procurement is well timed.

In order for the data resource to support a specific research community, the metadata describing the thematic nature of the data will need to be visible to the researcher, as well as themed catalogues and registries. This will also help to ensure the visibility and accessibility of the resource across a wider user base. The current EOSC Interoperability Framework supports the exchange of metadata describing a resource via a set of standardised APIs. Integration of the EOSC catalogues with future constructs will further support of digital research resources. For example, in the case where conceptual-themed research dataspaces are formalised and shared by means of the development of the appropriate framework which supports them.

The procurement team in WP8 could well employ the concept of *resource aggregators* for the distribution of data resources in a very similar way to that driving the mechanism demonstrating the distribution of commercial cloud based digital services (as resources) via the EOSC in the first procurement call (September 2022). These *resource aggregators* would be entities that were already registering other resource on EOSC catalogues, acting within the bounds of the EOSC RoP (Rules of Participation).

Note, by agreement with the EC, this document will evolve towards a formal procurement plan for commercial data in support of research dataspaces. This evolution will be based on regular input from the various stakeholders as an understanding of the nature of commercial data supporting research is unpacked and better understood.

2 Introduction

It is well understood that European researchers consume information from global industry resources focused on specific research domains, trends, and topics. This information is based on commercial activities and constitutes curated data that has been created, collected, developed, and published for commercial use, based on market demand.

Where this information contributes toward data-driven research outcomes, it is necessary for researchers to pay for the consumption of such. This could result from individual subscriptions to commercial journals, institutes or university libraries negotiating access to publications or datasets, or the aggregated use of commercially distributed data by various research support entities.

Task 8.1 in the EOSC Future project will endeavour to identify these data resources by working closely with researchers in one of the specific science disciplines to best understand their dependencies, and routes to, the consumption of such resources. Once we have identified commercial datasets/publications that are consumed by a significant segment of the researchers in a single domain, we will work to understand the model used by the entity distributing the resource in terms of procurement.

The team may well engage with an *aggregator*, in the form of a research support infrastructure, to distribute the data resource via the registration of such on the EOSC catalogues. This would require a mechanism by which the EOSC Future funding team may need to pre-procure a single purpose voucher from the data provider on behalf of the *aggregator* in a very similar way to which the team is procuring commercial digital services in the first procurement call on the project.

The *aggregator* would then register the data resource in the EOSC marketplace, focusing on the catalogue metadata describing the resource to ensure that the thematic nature of the data was made visible, not only to researchers, but also to other thematic data frameworks by means of standardised APIs. These would include the interoperability frameworks planned for the support of the European research dataspaces.

3 Research engagement and requirements analysis

3.1 Single discipline Proof of Concept (PoC)

The Task 8.1 team is considering Life Sciences as the first thematic group of researchers to reach out to in order to identify commercial data resources that are commonly consumed in support of research activities/outcomes. The EOSC Life project will be key to unpacking the data resources used by this community and singling out those that require subscription to, or procurement. As the EOSC Life project collaborates in the EOSC Future project, the team has already contacted the related team and have already presented on potential procurement activities to the clusters.

EOSC Life bridges the 13 Life Sciences research infrastructures from a data perspective. It is the project's ambition to 'to create an open, digital and collaborative space for biological and medical research'. 'EOSC-Life will make data resources from LS RIs 'FAIR' and publish them on the EOSC marketplace following guidelines and standards. Overall, this will drive the evolution of the RI repository infrastructure for EOSC and integration of the LS RI repositories.'

Commercial data consumed by researchers within this discipline would provide a comprehensive space within which the project can drive further relevance. In the registration of the components of the collaborative 'data space' on the EOSC marketplace through activities on the EOSC Future project, thematic data as it relates to the science discipline can be made visible to the researchers. Providing that the thematic nature of this data is clearly described during registration, it can be identified as a data resource for inclusion into a more formal science data space as related development evolves.

The identification of the commercial data described, and all related providers, is the first task of WP8.1, and as such will be unpacked in the evolution of this document.

3.2 The EOSC Rules of Participation

Key to the appropriate distribution of commercial resources in EOSC will be the provider rules of participation (RoP). This will ensure that aggregated demand for commercial data resources will result in an ethical, sustainable distribution of such in a way that is transparent and supports the general intent behind the way in which the EOSC will support European research activities and outcomes.

The EOSC Focus project presents additional opportunities to engage with the research community and focus on the delivery of pre-procured data resources via the EOSC Marketplace. The project is committed to the follow responsibilities:

- Provide an effective stakeholder forum;
- Consolidate and enhance existing monitoring frameworks;
- Identify strategic gaps to inform future integrations of the SRIA;
- Develop and test resourcing models for a sustainable EOSC;
- Implement the EOSC Rules of Participation (RoP);
- Collaborate with EOSC projects, other partnerships and international initiatives.

Early conversations with EOSC Focus suggest that the project could well assist in the facilitation of a discipline focused requirements gathering. Following the thematic approach, the Task 8.1 team proposes to work closely with the EOSC Focus project in engaging the Life Sciences community on commercial data resources currently consumed, and the mechanisms supporting these resources.

4 Procurement plan

4.1 Analysis of procurement options and feasibility study

GÉANT, based in Amsterdam the Netherlands, leads fully EU-rules-compliant tendering procedures to seek commercial providers on behalf of the European research and education community, supported by all its member National Research and Education Networks (NRENs). All procurements will be carried out in full accordance with the Dutch Public Procurement Act, governed by European Tendering laws and legislation, in which GÉANT acted as a central purchasing body.

A scheduled data resource procurement call (Q1 2023) run by the team could mimic the service resource call that will be finalised in Q4 2022, in that the commercial providers will reach out to aggregators (currently distributing non-commercial data through open-research activities in the EOSC environment) and collaborate on proposals for the distribution of commercial data resources via resource registration on the EOSC (Future) catalogues.

Due to the conceptual nature of the term 'data space' and the immaturity in terms of the development of any related research constructs, the procurement mechanism demonstrated will have to reflect a great deal of flexibility regarding the potential integration with more formal constructs in the future. Any pre-procured data resource that is registered as such on the EOSC Marketplace, will need to do so in a way that clearly identifies the target research audience. It this way, specific EOSC metadata describing the resource should already position it in the correct conceptual data space.

As with all pre-procured commercial resources (both services and data) available through EOSC constructs, the distribution mechanisms demonstrated by Task 8.1 activities will have to consider future procurements of these resources in terms of sustainable funding models. Consumable virtualised digital resources have driven the need for organisations across all sectors to transfer financial support for digital services from legacy projects requiring capital expenditure to models embracing the delivery of these resources in terms of ongoing operational costs. The way in which European projects focused on the 'on premises' development and hosting of digital resources have been funded for many years will need to evolve towards a more sustainable model of service delivery, if the European research community is to benefit from the unprecedented rate of innovation in the commercial sector

This section of the report will be developed once the demand has been defined as well as the providers of all significant commercial data resources (possibly currently consumed by the Life Sciences community as engaged via EOSC Focus) have been identified and consulted. The team would need to fully understand status quo in terms demand and distribution (and funding of these resources) before compiling the relevant EOSC Future procurement plan and related adoption calls.

5 Distribution of data resources

5.1 Definition and identification of aggregators for commercial data resource distribution

The introduction of the commercial services resource aggregator to the cloud services call (first T8.1 call to close on December 15th, 2022) can be applied in a very similar way to the distribution of commercial (and other) data resources via the EOSC Marketplace.

Several European research (e-) infrastructures, both internal and external to the EOSC Future project, have defined and taken responsibility for the distribution of data resources via the EOSC interoperability framework. These are currently primarily focused on the distribution of open, public data resources, although, if commercial data was pre-procured, there is no reason why these resources could not be catalogued on the marketplace in exactly the same way.

A data aggregator would need to demonstrate a mechanism by which the consumption of commercial data resources was carefully tracked so that 'top up' procurements could replenish available subscriptions once the initial adoption funding was exhausted. There could also be a mechanism by which a researcher/institute could procure subscriptions via the EOSC marketplace using their own grant funding if these were no longer available 'free at the point of use'.

Again, the metadata linking the commercial data resource to a conceptual research data space would need to be carefully applied when registering such a resource on the EOSC Marketplace. This would apply to both commercial and non-commercial data resources. This would ensure not only visibility regarding the science discipline, but also allow other formal research catalogues, not directly associated with the EOSC, to consume resource definitions and availability via the standardised EOSC interoperability APIs.

5.2 Current distribution mechanisms

There are several entities that are currently providing access to open data and content resources via the EOSC marketplace. Related services; this includes those in support of content distribution and tracking; data management tools that provide high accuracy data anonymisation and data management planning; metadata validation; data gateway creation… to name a few.

It is highly likely that these entities would be motivated to extend their role and relevance in the EOSC, and act as aggregators for the distribution of commercially curated data resources. T8.1 will identify these entities through various EOSC channels and investigate their appetite to participate in this way.

6 The European Data Spaces and EOSC

6.1 GÉANT working document on data spaces

There is currently a significant amount of work being done by the Commission to define and support the European research data spaces in terms of structure and framework. The output published thus far has been used as a basis for the article (developed by GÉANT) referenced in Appendix A.

Please refer Appendix A - 6.1 **Current state of play of the Common European Data Spaces & GÉANT / NRENs from a policy perspective**

7 Conclusions

The Task 8.1 team will endeavour to identify commercially produced, collated, and curated data sets that are consumed at scale by the European research community. As suggested, related efforts should possibly focus on specific science disciplines initially as targeted engagement of researchers via thematic projects and cluster projects will be easily accomplished within broad the EOSC Future stakeholder base.

Data resources identified in the analysis phase will clearly identify the commercial distribution channels for these resources. This will allow the task direct engagement with the providers and access to a good understanding of the underlying commercial models by which the data resources are current distributed. This will provide a basis for the procurement plan to be unpacked and detailed in the next iteration of this document.

The Task 8.1 team will ensure that all registration of data resources reflect the thematic nature of the data resource in the surrounding metadata. The standardised APIs delivered by the EOSC Future interoperability framework will then support these resources reflected in the catalogues of future catalogues which, in turn, will support researcher access to relevant data.

It will be important that the development of a framework specifically for the support of the European research data spaces will be in some level of collaboration with the European Open Science Cloud. This will ensure that all discipline specific data resources (not necessarily commercial) that have been/will be registered on the EOSC Marketplace will also be reflected in formal European data space catalogues.

8 Appendix A – Current state of play of the Common European Data Spaces & GÉANT / NRENs from a policy perspective

From a policy perspective, 2022 has so far contained only two key dates regarding the Common European Data Spaces or CEDS overall. The first was on the 23rd of February when the European Commission staff working document on the CEDS was released.[1] The purpose of the document serves both as an *aide-mémoire* and to present the status and future plans for the CEDSs, both to EC staff but also the public. The second important date was the announcement of the planned procurement of Simpl, the 'the smart middleware that will enable cloud-to-edge federations and support all major data initiatives funded by the European Commission, such as common European data spaces'[2], on the 30 May 2022. The first procurement is expected to be launched by the EC this month.

Why are the CEDSs being developed?

The EC has plans to bolster and grow the European economy, and digitisation is a key element of that. As the single market has been realised, so now is the beginning of the realisation of the **Digital Single Market**. A conceptual 'Schengen for data', the EC is looking to pool both public and private data within Europe more effectively. As data is being viewed as less of an output and more of a resource with value, the EC is looking to enhance the hybrid data created in Europe in order to stimulate the European economy at large. It hopes that this vision will be enabled via the creation of the CEDS, that are thematically grouped in different 10 areas: the Green Deal, Health, Energy, Manufacturing, etc. The European Open Science Cloud (EOSC) is also classed as one of these spaces. The groundwork for setting the rules as to how the CEDS will be developed was made first by the European strategy for data and subsequent first Data Governance Act[3] and secondly Data Act[4] (released on the same day as the SWD).

What is a CEDS?

It is important to note that a 'data space' is a loose term, which has been around in both policy and engineering circles well before the conception and legal beginnings of the CEDSs. BVDA has identified at least 10 different definitions of what a data space is[5]. The simplest definition the author can find is 'a platform for data sharing.' As mentioned above, the spaces will pool the input of both public and private data in order to create a hybrid data market in thematic areas. How the monetary value will be assigned to data sets, or how the monetary exchange for data sets will work in practice, is yet to be defined.

For information regarding business modelling, the best places to consult are i) those who have lessons learned from a procurement perspective with regards to the OCRE project and the ii) outputs from the Sustainability Working Group of the EOSC Association. For the most up to date information regarding data transfer from an ethical perspective, then information in this area will be most advanced coming from i) the European Health Data Space and ii) outputs from the Rules of Participation Working Group of the EOSC Association.

What will the CEDSs do?

According to the European strategy for data, the data spaces will include:

- the deployment of data sharing tools and services for the pooling, processing and sharing of data by an open number of organisations, as well as the federation of energy-efficient and trustworthy cloud capacities and related services;
- data governance structures, compatible with relevant EU legislation, which determine, in a transparent and fair way, the rights of access to and processing of the data;
- improving the availability, quality and interoperability of data – both in domain specific settings and across sectors.

[1] https://digital-strategy.ec.europa.eu/en/library/staff-working-document-data-spaces
[2] https://digital-strategy.ec.europa.eu/en/news/simpl-cloud-edge-federations-and-data-spaces-made-simple
[3] https://digital-strategy.ec.europa.eu/en/policies/data-governance-act
[4] https://digital-strategy.ec.europa.eu/en/policies/data-act
[5] Page 4, https://link.springer.com/content/pdf/10.1007/978-3-030-98636-0.pdf

Below we have a more technical illustration of how the group of spaces are situated in the wider data ecosystem (note in this image the mistaken omission of EOSC from what is 10 CEDSs).[6] It is important to note that the majority of CEDS are not being designed through the eyes of policymakers who are interested in R&E development. The spaces are primarily designed to strengthen the European data marketplace.

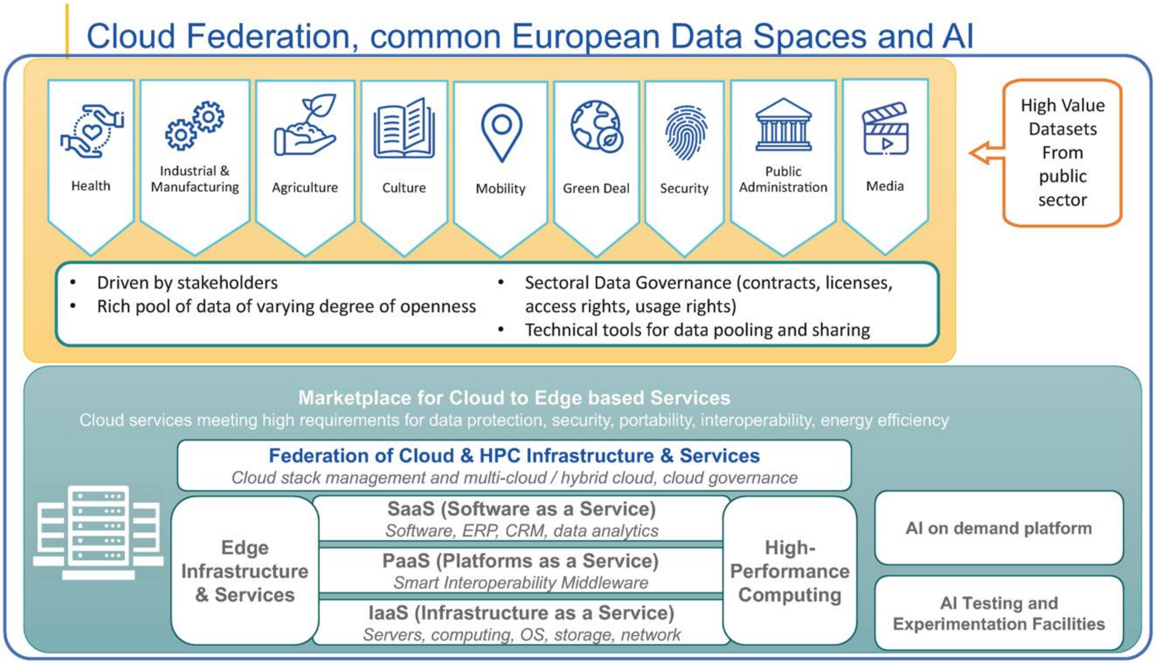

Cloud Federation, common European Data Spaces and AI
Health
Industrial & Manufacturing
Agriculture
Culture
Mobility
Green Deal
Security
Public Administration
Media
High Value Datasets From public sector
• Driven by stakeholders
• Rich pool of data of varying degree of openness
• Sectoral Data Governance (contracts, licenses, access rights, usage rights)
• Technical tools for data pooling and sharing
Marketplace for Cloud to Edge based Services
Cloud services meeting high requirements for data protection, security, portability, interoperability, energy efficiency
Federation of Cloud & HPC Infrastructure & Services
Cloud stack management and multi-cloud / hybrid cloud, cloud governance
Edge Infrastructure & Services
SaaS (Software as a Service)
Software, ERP, CRM, data analytics
PaaS (Platforms as a Service)
Smart Interoperability Middleware
IaaS (Infrastructure as a Service)
Servers, computing, OS, storage, network
High-Performance Computing
AI on demand platform
AI Testing and Experimentation Facilities

Figure 8.1: Overview of cloud federation, common European Data Spaces, and AI

What is the role of the GÉANT and NRENs towards the CEDSs?

GÉANT and NREN representatives are most acquainted with the work CEDSs will do from experience gained in both the EOSC environment and the OCRE project (with the latter especially from a procurement point of view). In addition to WP8 in EOSC future, there is WP2 activity dealing with liaison towards new strategic partners e.g. EuroHPC, GAIA-X etc., and this also includes an existing engagement strategy on how EOSC Future will engage with CEDS as and when they are set up.

Any draft actions proposed in this area should not look to contradict the planned model for Simpl, the middleware to be procured that will be:

- Anchored to specific use cases, covering a broad range of cases from sectoral data spaces (e.g. Agriculture, Genomics, Energy, Mobility) to Destination Earth, and from AI-on-demand to the European Open Science Cloud. Simpl will ensure that data sets and their infrastructures can be seamlessly interconnected and made interoperable.
- Smart and modular, to allow the replacement or addition of components without affecting the rest of the system, something that we hope to continue developing over time.
- Open source, allowing insights into all parts of the architecture (without any proprietary claims) and simple deployment.
- Green, scalable and elastic, by allowing a monitoring of its environmental performance, and the addition of new users without affecting performance.
- Secure and interoperable, where trust, confidence and compliance with regulations are built into the system. This implies an effortless sharing of resources between participants, regardless of their data processing environment. It creates an abstraction layer that enables data to flow across multiple providers and Member States.

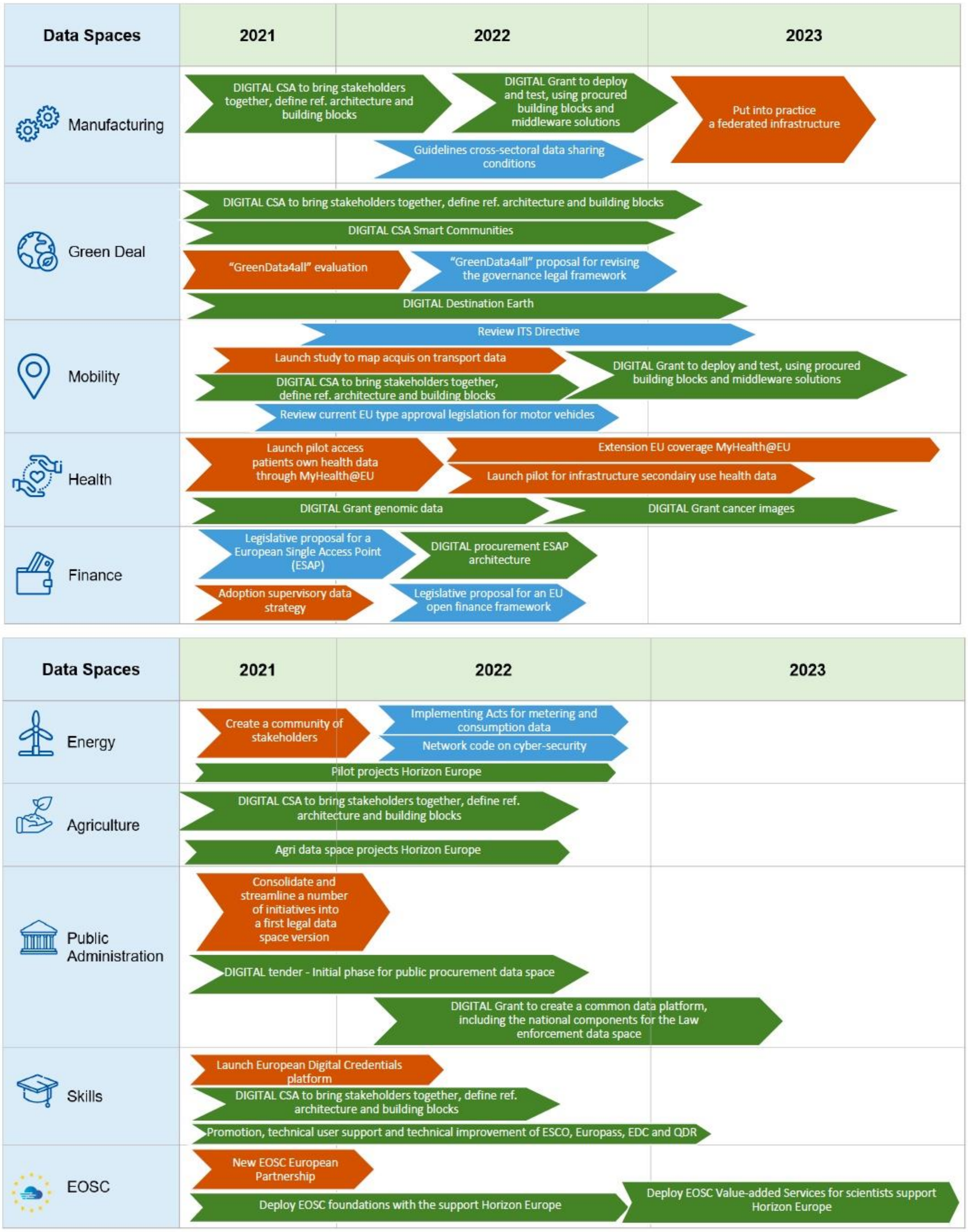
Conversations are being held both within NRENs and at the GÉANT level as how to best move forward with both potential service provision to the CEDSs when they become operational (where and if relevant), how to interact

[6] https://link.springer.com/chapter/10.1007/978-3-030-98636-0_16/figures/1

with the governance of different CEDSs (this is already happening with GÉANT and the NRENs in the European Health Data Space, for example) and will also scrutinize the first launch of the Simpl procurement.

What is the current status of each CEDS?

Please find below the timeline per CEDS, lifted from the Staff Working Document on the CEDS.



Data Spaces | 2021 | 2022 | 2023

Manufacturing
- DIGITAL CSA to bring stakeholders together, define ref. architecture and building blocks
- DIGITAL Grant to deploy and test, using procured building blocks and middleware solutions
- Put into practice a federated infrastructure
- Guidelines cross-sectoral data sharing conditions

Green Deal
- DIGITAL CSA to bring stakeholders together, define ref. architecture and building blocks
- DIGITAL CSA Smart Communities
- "GreenData4all" evaluation
- "GreenData4all" proposal for revising the governance legal framework
- DIGITAL Destination Earth

Mobility
- Review ITS Directive
- Launch study to map acquis on transport data
- DIGITAL Grant to deploy and test, using procured building blocks and middleware solutions
- DIGITAL CSA to bring stakeholders together, define ref. architecture and building blocks
- Review current EU type approval legislation for motor vehicles

Health
- Launch pilot access patients own health data through MyHealth@EU
- Extension EU coverage MyHealth@EU
- Launch pilot for infrastructure secondairy use health data
- DIGITAL Grant genomic data
- DIGITAL Grant cancer images

Finance
- Legislative proposal for a European Single Access Point (ESAP)
- DIGITAL procurement ESAP architecture
- Adoption supervisory data strategy
- Legislative proposal for an EU open finance framework

Data Spaces | 2021 | 2022 | 2023

Energy
- Create a community of stakeholders
- Implementing Acts for metering and consumption data
- Network code on cyber-security
- Pilot projects Horizon Europe

Agriculture
- DIGITAL CSA to bring stakeholders together, define ref. architecture and building blocks
- Agri data space projects Horizon Europe

Public Administration
- Consolidate and streamline a number of initiatives into a first legal data space version
- DIGITAL tender - Initial phase for public procurement data space
- DIGITAL Grant to create a common data platform, including the national components for the Law enforcement data space

Skills
- Launch European Digital Credentials platform
- DIGITAL CSA to bring stakeholders together, define ref. architecture and building blocks
- Promotion, technical user support and technical improvement of ESCO, Europass, EDC and QDR

EOSC
- New EOSC European Partnership
- Deploy EOSC foundations with the support Horizon Europe
- Deploy EOSC Value-added Services for scientists support Horizon Europe

Figure 8.2: Timeline per CEDS, lifted from the Staff Working Document on the CEDS

References

[1] EOSC Marketplace. 2022. [online] Available at: https://marketplace.eosc-portal.eu/

[2] SEE FOOTNOTE 5 Curry et al.