Grant Agreement N°: 101083927
Topic: DIGITAL-2021-CLOUD-AI-01
Type of action: CSA

Green Deal Data Space Foundation and its Community of Practice

D3.1: Initial Blueprint of the GDDS Reference Architecture

Deliverable no	3.1
Work package	WP 3 Technical Blueprint
Dissemination level	Public (PU)
Due data of deliverable	31 August 2023
Actual submission date	31 August 2023

Author(s)			
Partner	First Name	Last name	Email
CNR	Mattia	Santoro	mattia.santoro@cnr.it
CNR	Paolo	Mazzetti	paolo.mazzetti@cnr.it

Contributor(s)			
Partner	First Name	Last name	Email
UU	Kor	de Jong	k.dejong1@uu.nl
EOSC	Christian	Briese	Christian.Briese@eodc.eu
EGI	Marta	Gutierrez	marta.gutierrez@egi.eu
SURF	Paul	Gondim van Dongen	paul.gondimvandongen@surf.nl
ECMWF	Sebastien	Denvil	Sebastien.Denvil@ecmwf.int

Version	Date	Released by	Comments	Document status
0	19/04/2023	CNR		ToC
1	01/06/2023	CNR		Initial version
2	28/07/2023	CNR		Version for internal review
3	14/08/2023	CNR		Final version

Disclaimer

The information provided in this deliverable reflects the opinion of the authors and the GREAT (Green Deal Data Space Foundation and its Community of Practice) – project consortium under EC grant agreement 101083927and does not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein. It is important to note that the contents of this document have not been reviewed, endorsed, or approved by the European Commission (EC).

Executive Summary

On February 2020, the European Commission (EC) published a Communication introducing “A European strategy for data” (ESD) for the creation of “a single European data space – a genuine single market for data, open to data from across the world”. The strategy to achieve this vision is articulated around four main pillars: (i) a cross-sectoral governance framework for data access and use; (ii) investments in data and strengthening Europe’s capabilities and infrastructures for hosting, processing, and using data, interoperability; (iii) empowering individuals, investing in skills and in SMEs; and (iv) Common European data spaces in strategic sectors and domains of public interest. According to the ESD, Data Spaces should foster an ecosystem (of companies, civil society, and individuals) creating new products and services based on more accessible data. In addition, what distinguishes the Common European Data Spaces from other data sharing initiatives is its focus on preserving European values, balancing the flow and wide use of data, while preserving high privacy, security, safety, and ethical standards. One of the nine proposed sectorial European data spaces was the Green Deal Data Space (GDDS), for which the GREAT project is charged with developing an implementation roadmap, including a technical blueprint, governance scheme and priority datasets. This document describes the initial technical blueprint of the GDDS reference architecture.

As one of the key enablers for the green and digital transitions envisioned by the European Commission, the GDDS must be designed with a long-term perspective in mind. This means that the GDDS must be able to react and adapt to changes (e.g., the new functionalities and requirements stemming from such changes), particularly in the science/policy and technology contexts.

The design of the GDDS is based on the concept of Digital Ecosystem (DE). DEs focus on a holistic view of diverse and autonomous entities (i.e., the many heterogeneous and autonomous online systems, infrastructures, and platforms that constitute the bedrock of a digitally transformed society) which share a common environment. In search of their own benefit, such entities interact and evolve, developing new competitive or collaborative strategies, and, in the meantime, modifying the environment. What makes the ecosystem paradigm so powerful for adaptation to changes is that it focuses on the overarching values, i.e., the provisioning of ecosystem services. The preservation and enhancement of the ecosystem services are the driving factors of a digital ecosystem design. The belonging species are not subject to any predefined behavior, as long as this is not disruptive for the ecosystem services. Indeed, the inner structure of the digital ecosystem is free to vary over time adapting to contextual changes to preserve and enhance the ecosystem services.

The GDDS domain is characterized by a high level of heterogeneity, with many already existing data sharing initiatives that offer their resources to diverse consumers, which mirrors the current state of (geospatial) data sharing globally. Rather than assuming this situation will change, the digital ecosystem approach acknowledges and supports it. As a result, the GDDS DE must be established as a ‘soft’ infrastructure, a loosely federated

system based on minimal agreement for openness - i.e., the description and documentation of adopted specifications. Since establishing a single "common format" is not possible in a multidisciplinary context like GDDS, the challenge is how to transform a collection of disparate systems that use different technical standards into a digital ecosystem. This requires a minimal set of logical components that enable the ecosystem's digital environment. Thus, the GDDS DE soft infrastructure is comprised of two elements: (i) agreements (including technical standards) - these pertain to the governance sphere, which identifies the rules for participating in the GDDS DE; and (ii) minimal set of (logical) components creating the digital environment - these components are in charge of providing the required interoperability solutions to connect the data consumers and data sources participating in the GDDS DE. This technical blueprint identifies and describes the set of minimal components which will enable the provisioning of the GDDS DE Ecosystem Service (i.e., its high-level capability), defined as it follows: Secure, trusted and seamless sharing (i.e., discovery, access and use) of data to support Green Deal applications.

Finally, the document analyses the presented GDDS DE technical blueprint with respect to the Data Space Support Centre (DSSC) vision and other relevant initiatives (e.g., DSBA framework, DestinE, SIMPL, etc.) and provides inputs to the definition of the GDDS implementation roadmap.

Table of Contents

EXECUTIVE SUMMARY..4
1 INTRODUCTION..10
1.1 THE EUROPEAN STRATEGY FOR DATA..10
1.2 THE EUROPEAN GREEN DEAL..11
1.3 THE GREEN DEAL DATA SPACE..11
2 RATIONALE AND MAIN CONCEPTS..12
2.1 COMMON EUROPEAN DATA SPACES..13
2.2 TOWARDS A GREEN DEAL DATA SPACE..14
3 GREEN DEAL DATA SPACE AS A GEOSPATIAL DIGITAL ECOSYSTEM..19
3.1 GEOSPATIAL DIGITAL ECOSYSTEMS..19
3.2 GREEN DEAL DATA SPACE AS A DE..22
3.3 GDDS DESIGN METHODOLOGY..23
3.4 SOFT INFRASTRUCTURE..23
3.5 GOVERNANCE CHALLENGES..24
4 TECHNICAL BLUEPRINT OF GREEN DEAL DATA SPACE..25
4.1 ORTHOGONALITY OF DATA-SHARING AND SECURITY ARCHITECTURES..27
4.2 ARCHITECTURE DESCRIPTION..27
4.3 ENTERPRISE VIEWPOINT..28
4.4 INFORMATION VIEWPOINT..32
4.5 COMPUTATIONAL VIEWPOINT..35
4.6 ENGINEERING VIEWPOINT..43
4.7 TECHNOLOGY VIEWPOINT..46
4.8 SECURITY AND TRUST ARCHITECTURE..48
5 GDDS DE INTEROPERABILITY WITH OTHER INITIATIVES..53
5.1 DATA SPACE SUPPORT CENTRE TECHNICAL BLUEPRINT..54
5.2 DESTINATION EARTH DATA LAKE..57
5.3 SIMPL- SMART MIDDLEWARE PLATFORM (SMP)..59
5.4 DATA SPACE BUSINESS ALLIANCE – TECHNICAL CONVERGENCE..59
5.5 JRC REPORT ON EUROPEAN DATA SPACES..60
6 CONCLUSIONS AND INPUTS TO ROADMAP..64
7 REFERENCES..66
ANNEX A: POSSIBLE SERVICE INTERFACES AND DATA/METADATA MODELS TO BE SUPPORTED IN GDDS DE..70
ANNEX B: LEGAL AND ETHICAL ASSESSMENT METHODOLOGY..74

List of Figures

Figure 1 – The GDDS DE Soft Infrastructure 24
Figure 2 - Main actors involved in the creation, enhancement, and growth of the GDDS DE (UML Use Case diagram) 30
Figure 3 - Persistent and Unique Identifier of GDDS DE Logical Resource 33
Figure 4 - Core and Facilitators 36
Figure 5 - Modelling of Data Source Component 36
Figure 6 - Modelling of Data Consumer Component 37
Figure 7 – UML Diagram of Initial Set of Core Logical Components of the GDDS DE 39
Figure 8 - UML Diagram of Initial Set of Facilitators Logical Components of GDDS DE 41
Figure 9 - Example of Core Components Engineering Diagram 44
Figure 10 - Example of Facilitators Components Engineering Diagram 45
Figure 11 - Logical Architecture of GDDS DE Security Architecture 52
Figure 12 – Logical Components of the Security and Trust Architecture of the GDDS DE 53
Figure 13 - Technical Building Blocks from DSSC Taxonomy Document 54
Figure 14 - DEDL System Context (source: DestinE - System Framework - Data Lake - High Level Description & Architecture) 58
Figure 15 - Possible Development Roadmap of the GDDS DE 64

List of Tables

Table 1- High-level Functional/Non-functional Data Sharing Architecture Requirements 30
Table 2 - List of Initial Set of Core Logical Components... 39
Table 3 - List of Initial Set of Facilitators Logical Components 43
Table 4 - High-level Functional/Non-functional Security Architecture Requirements 50
Table 5 -Analysis of High-level Requirements from JRC Science for Policy Report on European Data Spaces.. 60
Table 6 - List of possible relevant service interfaces and metadata/data models for discovery and access 70

Glossary of terms

DEP	Digital Europe Programme
DS	Data Space
DSBA	Data Space Business Alliance
DSSC	Data Space Support Centre
DE	Digital Ecosystem
EC	European Commission
EGD	European Green Deal
EGDS	European Green Deal Data Space
EU	European Union
GREAT	Short Name to refer to Green Deal Data Space Foundation and its Community of Practice
HE	Horizon Europe
KPI	Key Performance Indicator
PPT	PowerPoint Template
TF	Task Force
WP	Work Package

1 Introduction

1.1 The European Strategy for Data

Fair and trustworthy data sharing that can support the development of innovative products and services is at the core of the European Commission's agenda [1]. The priority "A Europe fit for the digital age" [2] guides the European Commission's policy agenda for the period of 2019-2024, culminating in the EC's vision for Europe's digital transformation "2030 Digital Compass: the European way for the Digital Decade" [3] which sets ambitious targets aimed at strengthening digital sovereignty through specific actions on data, technology and infrastructures. The Annual Single Market Report [4], published in 2023, marks the 30th anniversary of the Single Market, and highlights the ambition to create a single EU data economy through a data-driven Single Market where interoperability within and across data spaces is ensured [1].

On February 2020, the European Commission (EC) published a Communication introducing "A European strategy for data" (ESD) for the creation of "a single European data space – a genuine single market for data, open to data from across the world" [5]. The strategy to achieve this vision is articulated around four main pillars:

- A cross-sectoral governance framework for data access and use;
- Enablers: Investments in data and strengthening Europe's capabilities and infrastructures for hosting, processing, and using data, interoperability;
- Competences: Empowering individuals, investing in skills and in SMEs;
- Common European data spaces in strategic sectors and domains of public interest.

According to the ESD, Data Spaces should foster an ecosystem (of companies, civil society, and individuals) creating new products and services based on more accessible data. In addition, what distinguishes the Common European Data Spaces from other data sharing initiatives is its focus on preserving European values, balancing the flow and wide use of data, while preserving high privacy, security, safety, and ethical standards. One of the nine proposed sectorial European data spaces was the Green Deal Data Space, for which the GREAT project is charged with developing an implementation roadmap, including a technical blueprint, governance scheme and priority datasets.

To support the ESD, in November 2020, the EC proposed a Data Governance Act [6] aiming at increasing trust in data sharing and facilitating data reuse. In February 2022, the EC proposed a Data Act [7] to make more data available for use in line with EU rules and values. The Data Governance Act creates the processes and structures to facilitate data, while the Data Act clarifies who can create value from data and under which conditions. Finally, in the framework of the Open Data Directive [8], the European Commission adopted an Implementing Act [9] specifying certain "High Value Datasets" that public sector organizations will have to make available free of charge, in machine-readable format.

1.2 The European Green Deal

In parallel with the "digital transition" described above, there is an equally important "green transition". In December 2019, the European Commission unveiled its flagship action plan to tackle climate change, the European Green Deal. Through this strategy, the European Union (EU) aims to become the first resource-efficient and competitive economy without net emissions of greenhouse gases by 2050.

The European Green Deal has been expressed through several policies and plans, charting a comprehensive course for action, supported by a growing number of legislative and regulatory actions. The Green Deal sets ambitious objectives, including restoring degraded ecosystems at land and sea across Europe with the 2030 Biodiversity Strategy and reducing greenhouse gas emissions to zero by 2050 with the European Climate Law and the Zero Pollution Action Plan . In addition to regional action, part of the action plan is to increase the EU's "green diplomacy" and demonstrate EU leadership in multilateral fora to increase collective effort and reach the objectives of the Paris Agreement and the United Nations Sustainable Development Goals (UN SDGs).

Ambitious action plans like the European Green Deal require an abundance of resources, including viable data. Data allows governments to identify risks, tailor policy response and resource allocation, monitor progress and identify trends. However, serious data gaps remain in the global fight against climate change. While some consequences of climate change are irreversible, data gaps and analytics deficits need to be addressed.

1.3 The Green Deal Data Space

The Green Deal Data Space stands at the intersection of two major European policy initiatives: the EU Strategy for Data and the European Green Deal. The GDDS will be designed and implemented to exploit the potential of data to effectively support the Green Deal priority actions, empowering policy makers, businesses, researchers, and citizens, from Europe and around the world, to jointly tackle issues such as climate change, circular economy, zero pollution, biodiversity protection, deforestation and compliance assurance. Out of the many European Green Deal strategic actions, the GREAT project focusses on three priorities (Biodiversity 2030, Zero Pollution and Climate change), in order to effectively capture the diversity of requirements across the full range of the European Green Deal. These three initiatives are interlinked with other EGD strategic actions, approximate the full scope of the GDDS, as well as complementing actions being addressed by other thematic data spaces (such as the "Farm to Fork Strategy" being addressed by the agricultural data space).

2 Rationale and Main Concepts

The ESD recognizes the importance of data-driven innovation for the benefit of society, highlighting as one of the examples its application to the European Green Deal objectives.

> *Data-driven innovation will bring enormous benefits for citizens, for example through improved personalised medicine, new mobility and through its contribution to the European Green Deal. [European Strategy for Data] [5]*

Besides, what distinguishes the Common European Data Spaces from other data sharing initiatives is its focus on preserving European values.

> *In order to release Europe's potential we have to find our European way, balancing the flow and wide use of data, while preserving high privacy, security, safety and ethical standards. [European Strategy for Data] [5]*

Therefore, it can be recognized from the ESD that the two main technical challenges in building a Common European Data Space are data interoperability and security/trust.

Several EU horizontal programmes will support the development of common European data spaces through various funding actions [10], notably, the Digital Europe (DIGITAL) programme for digital deployment initiatives, the Horizon Europe (HORIZON) programme for research and innovation, the Connecting Europe Facility (CEF) for digital infrastructures, and the European Open Science Cloud (EOSC). Moreover, the recovery plans of several Member States also support actions on European data spaces [10].

Specifically, the DIGITAL Work Programme 2021-2022 [11] planned a set of dedicated calls for funding the preparatory phases of the European data spaces in the listed domains. The requested outcomes include the design of the overarching architecture – i.e., the technical blueprint - the description of the proposed governance, the identification of high priority datasets, and, as a final step, the definition of a roadmap for the implementation.

> *The GDDS will interconnect currently fragmented and dispersed data from various ecosystems. [11] [Digital Europe - Work Programme 2021-2022]*

The GREAT (The Green Deal Data Space Foundation and its Community of Practice) project was selected for the execution of the preparatory actions for the Green Deal Data Space (GDDS).

Define the technical blueprint of the GDDS reference architecture explaining how existing (and planned) data ecosystems (at European, national, regional, and local level) can be connected to provide an interoperable, secure data sharing environment which allows seamless discovery and use of available data. [GREAT Proposal]

2.1 Common European Data Spaces

According to a dedicated Commission Staff Working Document, a Common European Data Space "brings together relevant data infrastructures and governance frameworks in order to facilitate data pooling and sharing" [10]. The document also lists a set of key features for the Common European data spaces:

- A secure and privacy-preserving infrastructure to pool, access, share, process and use data.
- A clear and practical structure for access to and use of data in a fair, transparent, proportionate and non-discriminatory manner and clear and trustworthy data governance mechanisms.
- European rules and values, in particular personal data protection, consumer protection legislation and competition law, are fully respected.
- Data holders will have the possibility, in the data space, to grant access to or to share certain personal or non-personal data under their control.
- Data that is made available can be reused against compensation, including remuneration, or for free.
- Participation of an open number of organizations/individuals.

The document also provides more specificity on a couple of key technical aspects:

- Participants in common European data spaces will be encouraged to use the common technical infrastructure and building blocks which will allow the data spaces to be built in an efficient and coordinated manner.
- To avoid fragmentation, high integration costs and the creation of silos, the common European data spaces could develop on international standards, INSPIRE (for spatial data) and FAIR principles to favor interoperability, exploitation of data on EU computing infrastructures (e.g., cloud and HPC).

Recently, the "European Data Spaces - Scientific Insights into Data Sharing and Utilisation at Scale" report [1] was released by the EC JRC analyzing the main EU policy documents to identify a set of key principles and high-level requirements for the Common European Data Spaces.

The analysis carried out by the document highlights that, on one side, "from a technical perspective, a single architecture or stack of technologies and standards cannot be universally applied" [1]. However, the document also recognizes that "a minimum stack of protocols and specifications [...] is highly desirable" and the "forthcoming European Data Innovation Board, defined by the Data Governance Act and supported by the Data Spaces Support Centre, should play a central role in the choice of such technologies and standards" [1]. It is worth to note that the constant evolution of technologies will require an iterative refinement/review of such a selection.

2.2 Towards a Green Deal Data Space

2.2.1 An Ever-Changing Landscape

As one of the key enablers for the green and digital transitions envisioned by the European Commission, the GDDS must be designed with a long-term perspective in mind. This means that the GDDS must be able to react and adapt to changes (e.g., the new functionalities and requirements stemming from such changes), particularly in the science/policy and technology contexts. In fact, as outlined in the following two sub-sections, the last years were characterized by the occurrence of several changes both in the science/policy and in the technological context. Coping with such changes would be relatively easy if they happened in a predictable way, allowing to schedule periodical revisions and updates of the GDDS enabling infrastructure (including its components, software stack, etc.). Unfortunately, this is not the case: changes happen continuously, especially in the technological context, and the design of the GDDS must take this into account, to avoid the risk of early obsolescence.

Therefore, the real lesson learned from the past is that change is unpredictable but not unexpected. Therefore, the design and implementation of the GDDS must be flexible enough to accommodate the changes in the science, policy and technological contexts.

Changes in the Science/Policy Context

Changes in the Science/Policy context affect the design and development of solutions for data sharing and exploitation. An example is GEO, which in its Strategic Plan 2016-2025 focused the scope of GEO and GEOSS on targeting societal challenges, highlighting that "Earth observations are an indispensable component to measure and monitor our progress towards addressing societal challenges" [11]. To this aim, the GEO XIII Plenary in 2016 approved an Engagement Strategy [13] and selected three key policy priorities to guide GEO's efforts over the medium term: Paris Agreement on Climate Change, Sendai Framework on Disaster Risk Reduction, and the United Nations Agenda for Sustainable Development. At its 18th meeting in September 2020, the GEO Programme Board reviewed and endorsed a proposal from the Urban Resilience Subgroup recommending that Urban Resilience be recognized as a fourth GEO engagement priority [14].

The identification of engagement priorities and their change over time impacted on the GEOSS evolution shifting focus from data sharing to the generation of knowledge from EO, for example to compute indicators, such as the UN Sustainable Development Goals indicators. They also highlighted the role of decision-makers and policy-makers as end-users of GEOSS.

The focus on societal challenges made GEO exposed to changes in the Science/Policy context. This is reflected in the initial choice of Engagement Priorities (Paris Agreement on Climate Change, Sendai Framework on Disaster Risk Reduction, United Nations Agenda for Sustainable Development), with the later proposal of a fourth Engagement Priority (Urban Resilience) and uprising challenges suggested by occurring events such as the raised interest on the environmental impact on health following the pandemic of 2020. Such changes affected the GEOSS design and development due to different requirements concerning data (spatial-temporal resolution and coverage, uncertainty, etc.) and modelling.

Changes in the Technological Context

In the last decades several technologies affected - or had the potential to affect - the landscape of geospatial data sharing and processing. Just to mention a few of them:

- *Cloud technologies* allow to store big satellite data and to remotely process them. Cloud-based platforms were developed specifically for accessing, visualizing, and processing EO data (Google Earth Engine, Copernicus DIAS, etc.). They demonstrated how remote processing of data can be convenient (e.g., for performance, cost-effectiveness, etc.), thus suggesting a *mobile code* approach instead of the traditional *search and download* approach.
- *Data cubes*, pre-processing datasets during the ingestion phase, allow accessing the so-called Analysis Ready Data (ARD), potentially reducing the data preparation phase, and making processing for knowledge generation more efficient.
- The *Internet-of-Things* (IoT), enabling the networking of sensors and actuators, promising a new era of in-situ data acquisition but also potentially a new data deluge with new challenges on storing, accessing, and processing these new datasets.
- *Edge computing*, moving data (pre-)processing close to the sensors can help to address IoT challenges, reducing the required bandwidth, and envisioning a Cloud Continuum supporting data processing.
- *Artificial Intelligence* (AI) boosted by the advancement of data-driven approaches based on Machine Learning (ML) and Deep Learning (DL) promise to deeply affect EO data processing as and other fields.

Each of these new technologies is a potential enabler for new capabilities in the GDDS, accelerating and improving the digital transformation in the Green Deal sector. At the same time, however, they raise also challenges on, e.g., how to exploit/integrate such technologies in the GDDS and understanding how the resulting innovation affects (positively or negatively) the overall Green Deal-related digital environment.

2.2.2 Geospatial Data Interoperability Challenges

Green Deal-related data mainly belong to the geospatial information realm, that is “information concerning phenomena implicitly or explicitly associated with a location relative to the Earth” [15]. Geographic Information is represented and conveyed through (geo)spatial data that is “any data with a direct or indirect reference to a specific location or geographical area” [16].

The geoinformation world is characterized by great complexity with many actors involved, including:

- *Data producers* who acquire observations (e.g., through sensors);
- *Data providers* who distribute data managing data centres, long-term preservation archives, Spatial Data Infrastructures, etc.
- *Overarching initiatives* that influence the geoinformation world, designing new solutions, building disciplinary or interdisciplinary systems of systems, managing high-level expert groups, etc.
- *Technology providers* who develop and distribute technological solutions for geospatial data management and sharing
- *Cloud providers* who manage complex infrastructures on behalf of other actors such as data providers or application developers
- *Application developers* who make use of data to build applications for end-users
- *End-users* who utilize data

In such a context, interoperability is clearly perceived as one of the main issues, even considering only its technological facet. Indeed, actions and concerns of different actors have an impact in terms of technological choices.

- *Data producers* are mostly focused on data and metadata models and formats. Multiple standards have been defined addressing issues which are specific for different disciplinary domains, such as HDF, netCDF and GRIB for EO data, ESRI Shapefile or OGC GML for feature type information, and many others. Proprietary formats are still widespread.
- *Data providers* are mainly focused on data sharing services. As for data models and formats, several standards have been designed and adopted in different disciplinary domains. For example, in the biodiversity context TDWG standards are widely adopted, in the meteo-ocean community THREDDS Data Server is a widespread technology. OGC standard services are commonly adopted in the GIS community. Light specifications like KML (now an OGC standard) or OpenSearch are also common. OAI-PMH is a standard for long-term preservation archives.

- *Overarching initiatives* influence technological aspects in several ways, in particular on data management (e.g., the Data Management Plan guidelines in Horizon Europe programme), data harmonization (e.g., WMO information systems specifications) and data sharing, including policy (e.g., RDA).
- *Technology providers* contribute to the heterogeneity providing many different competing solutions for geospatial data sharing. While some of them have adoption of standards as an objective, others (often from big players) prefer to push their own proprietary solutions.
- *Cloud providers* affect technologies providing new data storage and processing capabilities requiring new solutions for integration with traditional systems.
- *Application developers* contribute to the heterogeneity of the geoinformation world because they provide geospatial applications adopting different technologies, from operating systems and related ecosystems (e.g., Linux, Microsoft, Apple, Google Android), to development platforms (e.g., Java, Python, Javascript) and libraries.

The interoperability issue is explicitly recognized also in the ESD: "data producers and users have identified significant interoperability issues which impede the combination of data from different sources within sectors" [5]. Unfortunately, as it will be explained later, the lack of agreed interoperability standards in the Green Deal sector is indeed an issue, but it is more the consequence of the complexity of the geospatial world than the reason of it. Including many actor categories, many disciplines, and many stakeholders (public authorities, private companies, citizens, etc.) the complexity of the geospatial world makes it impossible to agree on a single standard or even on a small set of standards and, later, impose and enforce its adoption [17] [18] [19] [1].

2.2.3 Security and Trust Challenges

Security challenges mainly stem from the fragmented nature of the Green Deal related data sharing infrastructures and initiatives.

a) Each provider of the GDDS must be able to define its own data policies and these must be supported at the GDDS level. Unfortunately, in the environmental domain, the efforts on policy harmonization are still limited, resulting in the need of supporting a highly heterogenous set of data policies, which makes access control complex and difficult to maintain.
b) Data policy enforcement can be implemented as an end-to-end solution or as a simpler access control mechanism. While the former has the advantage of enforcing data usage policy conformance even after the data has been transferred, it restricts the usage of data to an environment which supports the adopted end-to-end technological solution. A lighter access control mechanism, on the other end, removes such a restriction but leaves the respect of the data usage policy with the user, possibly lowering the trust by data providers.

Trust refers to ensuring that a claim (e.g., “the user with ID ‘id1’ is a non-commercial user”) is true. Achieving trust in a context like the GDDS can be built on top of two pillars:

a) Technical: to be able to ensure (verify) that the claim is from a certain organization.

b) Governance: acknowledge an organization as trustworthy, including the possibility of having different levels of trustworthiness for different types of claims.

At the technical level, there exists several solutions which provide the desired functionality. It is important to note here that compatibility with DSSC and, in turn, other sectorial Data Spaces is key to build an inter-Data Space trusted environment underpinning the envisioned single digital market.

2.2.4 Green Deal Related Initiatives

The creation of the Common European Data Spaces is also related to other parallel policy and technical initiatives stemming from the ESD. In particular, the GDDS, which aims at supporting the Green Deal priority actions, will leverage actions implemented in the ‘GreenData4All’ and ‘Destination Earth’ initiatives. The ‘GreenData4All’ initiative aims at evaluating and possibly reviewing the INSPIRE Directive, making it easier for EU public authorities, businesses, and citizens to support the transition to a greener and carbon-neutral economy, and reducing administrative burden. The Destination Earth (DestinE) initiative will bring together European scientific and industrial excellence to develop a very high precision digital model of the Earth (digital twin of the Earth) [10]. The objective of the DestinE initiative is therefore to deploy several highly accurate digital replicas of the Earth (Digital Twins) in order to monitor and simulate natural as well as human activities and their interactions, to develop and test “what-if” scenarios that would enable more sustainable developments and support European environmental policies. DestinE faces the challenge to manage and make accessible the sheer amount of data generated by the Digital Twins and observation data located at external sites. This data must be made available fast enough to support analysis scenarios by users of the DestinE Core Service Platform. Other relevant initiatives, described in the next paragraphs, are being developed with the intent to support the concept of Data Spaces and their implementation.

Simpl is the smart middleware that will enable cloud-to-edge federations and shall support all major data initiatives funded by the European Commission, such as common European data spaces[1]. The objective is to procure a large-scale modular and interoperable open-source smart European cloud-to-edge middleware platform. Such capability will allow the integration of data infrastructures and services that will address the needs of the different data spaces and enable the realisation of the European Cloud Federation.

[1] https://digital-strategy.ec.europa.eu/en/news/simpl-cloud-edge-federations-and-data-spaces-made-simple

In the private sector, the Big Data Value Association (BDVA), FIWARE Foundation, Gaia-X and the International Data Spaces Association (IDSA) decided to join forces and formed the Data Spaces Business Alliance (DSBA) aimed at driving the adoption of data spaces across Europe and beyond. Members of the DSBA agreed to work towards defining a common reference technology framework, based on the technical convergence of existing architectures and models, leveraging each other's efforts on specifications and implementations. The goal is to achieve interoperability and portability of solutions across data spaces, by harmonizing technology components and other elements.

Finally, it is worth mentioning large global initiatives like the Global Earth Observation System of Systems (GEOSS), developed by the Group on Earth Observation (GEO), a voluntary partnership of more than 100 national governments and 100 Participating Organisations. GEOSS was developed with the aim of achieving comprehensive, coordinated and sustained observations of the Earth and improve monitoring and prediction of the state of the planet. From the beginning GEOSS was conceived as a "system of systems", that is a loose confederation of existing and future Earth observation and data management systems. Because of the voluntary nature of GEO, the development of its system of systems has happened largely from the bottom-up exploiting opportunities and the willingness of the partnering organisations to contribute to this global endeavour. In the context of GEO, EuroGEO is the regional initiative which promotes cooperation at the European level. The GEO/EuroGEO is particularly relevant for the GDDS for several reasons. It provides an entry point to discover and access millions of datasets from about 200 data sources globally. GEOSS was built as a system of systems building on existing capacities. Finally, the GDDS can be a major contribution to GEO in the context of the EuroGEO initiative.

3 Green Deal Data Space as a Geospatial Digital Ecosystem

In this section we introduce the concept of Digital Ecosystem, along with some examples, and how it applies to the Geospatial world. Then we describe why such a paradigm fits the vision of the Green Deal Data Space. Finally, we introduce the design methodology for the GDDS as a Digital Ecosystem.

3.1 Geospatial Digital Ecosystems

A Geoscience Digital Ecosystem can be defined as a "system of systems that applies the digital ecosystem paradigm to model the complex collaborative and competitive social domain dealing with the generation of knowledge on the Earth planet" [20]. Inspired by this definition, we can consider a broader Geospatial Digital Ecosystem (GDE) as a *system of systems that applies the digital ecosystem paradigm to model the complex collaborative and competitive social domain dealing with the generation of knowledge from geospatial information*.

The Digital Ecosystem (DE) paradigm stems from the concept of natural ecosystems [21]. DEs focus on a holistic view of diverse and autonomous entities (i.e., the many heterogeneous and autonomous online systems, infrastructures, and platforms that constitute the bedrock of a digitally transformed society) which share a common environment. In search of their own benefit, such entities interact and evolve, developing new competitive or collaborative strategies, and, in the meantime, modifying the environment [22]. In the geospatial domain, DEs are called to enable the coevolution (i.e. the complex interplay between competitive and cooperative business strategies) of public and private organizations around the new opportunities and capacities offered by the digital transformation of society – Internet, big data, and computing virtualization processes represent some of the main engines of innovation, giving rise to an entirely new type of geospatial ecosystems [20].

A Natural Ecosystem can be characterized through its Ecosystem Functions and Services. Ecosystem Functions include the physicochemical and biological processes that occur within the ecosystem to maintain terrestrial life. Ecosystem services are the set of ecosystem functions that are directly linked to benefit human well-being. While the interaction among the species with the environment can vary, making the ecosystem adapt to external and internal changes, some of these changes can affect the Ecosystem Services and become disruptive. This is the reason why Natural Ecosystems need management and protection.

The same paradigm can be applied to the digital domain. In a Digital Ecosystem, diverse and autonomous entities – i.e., digital 'species' – share a common digital environment, and in search of their own benefit, they interact and evolve, developing new competitive or collaborative strategies, and, in the meantime, modifying the environment. Also, in the Digital Ecosystem it is possible to identify Ecosystem Functions, that are informational processes, and Ecosystem Services that are those functions of value for the Society.

What makes the ecosystem paradigm so powerful for adaptation to changes is that it focuses on the overarching values, i.e., the provisioning of ecosystem services. The preservation and enhancement of the ecosystem services are the driving factors of a digital ecosystem design. The belonging species are not subject to any predefined behavior, as long as this is not disruptive for the ecosystem services. Indeed, the inner structure of the digital ecosystem is free to vary over time adapting to contextual changes to preserve and enhance the ecosystem services. It is worth noting that the digital ecosystem does not require or expect that species deliberately work for the ecosystem. Instead, it accepts that they work for their own benefit while they contribute to the ecosystem functions and services. This works as far as species gain benefit from belonging to the ecosystem. Each species must find its own compromise between the desire of Autonomy (to be free to pursue its own benefit without any constraint) and the advantages of Belonging (to contribute to the ecosystem to gain an indirect benefit).

Adapting to contextual changes while preserving and enhancing ecosystem services is a key characteristic of digital ecosystems, which must be free to evolve to cope with such changes. However, during this evolutionary process, the values and services of the digital ecosystem must not be lost. Therefore, it is necessary to identify the essential characteristics of the ecosystem, associated with its services, that must be considered as the sole and real invariants. They are the immutable core that must not change under penalty of the destruction of the ecosystem – i.e., the loss of any ecosystem service.

Preserving the invariants requires a cybernetic mechanism of control and communication that is part of the digital ecosystem governance. For example, a governance process must be able to address possible conflicts such as belonging vs. autonomy – i.e., the possibility of conflict between participating system values and of overall ecosystem values.
Three main types of governance styles can be recognized for digital ecosystems:

- **Directed**: the ecosystem is centrally managed to ensure the long-term fulfillment of the ecosystem purposes, as well as any new purpose the system owners might wish to address.
- **Collaborative**: like in the directed ecosystems, there are recognized objectives, however, there is not any central authority, and the constituent systems collaborate to fulfill the agreed upon central purposes
- **Acknowledged**: As in the directed ecosystem, there is a central management organization, but the constituent systems maintain their autonomy only contributing to the (acknowledged) ecosystem purposes.

3.1.1 Examples of Digital Ecosystems

To evaluate the feasibility of an information sharing system as a (Geospatial) Digital Ecosystem it is first interesting to search for successful examples of Digital Ecosystems. A first clear example is the World Wide Web: it is built around a set of architectural principles – Identification, Interaction and Representation – and related technical specifications – mainly URL, HTTP, HTML, and their descendants. Currently the WWW is an ecosystem hosting a diversity of species, including institutions, organizations, companies, citizens. They have their own interests and values, but all of them limit their Autonomy using the WWW to publish and access information. They find that Belonging to the Web – i.e., accepting its governance and technological constraints – is acceptable because they get something in return – i.e., access to resources, visibility – that helps them to achieve their objectives – i.e., business opportunity, social interactions, etc. From their own point-of-view belonging to the Web is better than being fully autonomous. The WWW has many different functions, but it is valuable as an ecosystem supporting (unstructured) information sharing. Species have their own interest to pursue, but all of them contribute to the information sharing which can be considered as its ecosystem service. The WWW underwent deep changes since its birth in mid-90s. It was able to support new devices – e.g., mobile phones, web sensors –, new applications – e.g., search engines, social networks, e-commerce, e-governments -, new users - e.g., companies, public administrations, citizens. It is worth noting that none of them was anticipated in the design of the WWW which was designed as a ‘simple’ system for hypertext sharing.
There are other valuable examples of Digital Ecosystems. In recent years Software Ecosystems evolved as Business Ecosystems built around one or more core (software) technologies. Google, with Android, and Apple, with iOS, built examples of successful Ecosystems. Developing the Android operating system and opening it to external developers, Google created an ecosystem hosting several species. It also started a virtuous cycle with an increasing number of applications: more apps are available and more devices

are sold; more Android devices exist, and more developers are encouraged to create Android apps.

Of course, there is a fundamental difference between the WWW and the Android (or Apple) ecosystems. The WWW is not controlled by a single organization: its governance is distributed among different organizations, and the constituent systems maintain their autonomy i.e., it is an acknowledged ecosystem. On the other hand, typical software ecosystems like Google's Android or Apple's iOS are controlled by a single organization, the one that controls the core technologies, i.e., it is a directed ecosystem.

3.2 Green Deal Data Space as a DE

The GDE paradigm fits particularly well with the vision of the Common European data spaces, and, specifically, the GDDS. This is supported by two main characteristics of the GDDS:

- There are already existing (geospatial) data systems – and even limited ecosystems - managed by organizations according to their own mandate and governance. Due to their autonomy, they should not be considered simply as technological assets to leverage but as evolving digital "species" to host.
- There is no closed list of use-cases and related applications to build on the data space. It is anticipated that a data space will suggest and enable unexpected applications.

The first point is explicitly mentioned in the Commission Staff Working Document on Common European Data Spaces [10] and reiterated in the Green Deal Data Space call which asks for "a blueprint that connects existing national, regional and local data ecosystems [...] The Green Deal data space will interconnect currently fragmented and dispersed data from various ecosystems, both for/from the private and public sectors" [23].

The second point can be inductively supported. In the past, data sharing initiatives were designed without contemplating applications that are now considered important if not essential. For example, the first Web Geographical Information Systems (GIS) did not (and could not) consider mobile device limitations and hence app support; old data sharing infrastructures based on the *data search and download* pattern were not able to exploit cloud processing capabilities; more recently, Artificial Intelligence (AI) and Machine Learning (ML) applications and Digital Twins are challenging the existing data sharing and processing systems. Therefore, it is expected that a data space will need to evolve supporting applications that we cannot now imagine. A solution is to build a system made for changing, i.e., a digital ecosystem that is open to (non-disruptive) changes by design.

It is worth noting that the Position Paper on "Design Principles for Data Spaces" published by the OPEN DEI project[2] mentions the ecosystem nature of the data spaces defining them as "a federated data ecosystem within a certain application domain and based on

[2] https://www.opendei.eu/

shared policies and rules" [24]. However, it does not provide a definition of what it means by 'data ecosystem', which in this document we define as the above-described Geospatial Digital Ecosystem (GDE)

3.3 GDDS Design Methodology

Designing a digital ecosystem is not the same as designing a traditional information system. The latter aims at supporting a predefined set of intended use-cases with the provision of the best technical solution which complies with its requirements and constraints. Instead, the outline of a digital ecosystem should first identify the (high-level) ecosystem service to provide and then design a *satisficing*[3] architecture. Scenarios and use-cases are important but just to validate the architecture and they should be identified to cover a spectrum of potential applications as wide as possible. Changes have less impact in digital ecosystems than in traditional systems, because: a) they can introduce new use-cases and scenarios as far as they do not disrupt the ecosystem service; b) a satisficing architecture can easily accommodate changes remaining a satisficing architecture, while an optimal architecture can likely become suboptimal.

3.4 Soft Infrastructure

Another key concept for the effective design of a DE is that of "soft infrastructure". A soft infrastructure is invisible, made up of technology neutral agreements and standards, on how to participate in an ecosystem [24].

As outlined in previous sections, the GDDS is characterized by a high level of heterogeneity, with many already existing data sharing initiatives that offer their resources to diverse consumers, which mirrors the current state of (geospatial) data sharing globally. Rather than assuming this situation will change, the digital ecosystem approach acknowledges and supports it.

As a result, the GDDS DE must be established as a 'soft' infrastructure, a loosely federated system based on minimal agreement for openness - i.e., the description and documentation of adopted specifications. Establishing a single "common format" is not possible in a multidisciplinary context like GDDS. Just to give one example: data models for climatological studies require multidimensional time while data models for biodiversity applications require species taxonomies. A potential single standard would easily become too complex, posing an unacceptable high entry barrier for producers and consumers.

The challenge is how to transform a collection of disparate systems that use different technical standards into a digital ecosystem. The solution is not to eliminate fragmentation, but to conceal it as much as possible and necessary. This requires a minimal set of logical

[3] 'satisficing' is a term coined by the economist Herbert Simon to better model the behaviour of the 'rational agent' who typically does not search for an optimal decision which would require too much time and effort, but just stop their search at the first occurrence of a satisfying and sufficing solution.

components that enable the ecosystem's digital environment. Thus, the GDDS DE soft infrastructure (Figure 1) is comprised of the following two elements:

- Agreements (including technical standards): these pertain to the governance sphere, which identifies the rules for participating in the GDDS DE.
- Minimal set of (logical) components creating the digital environment: these components are in charge of providing the required interoperability solutions to connect the data consumers and data sources participating in the GDDS DE.

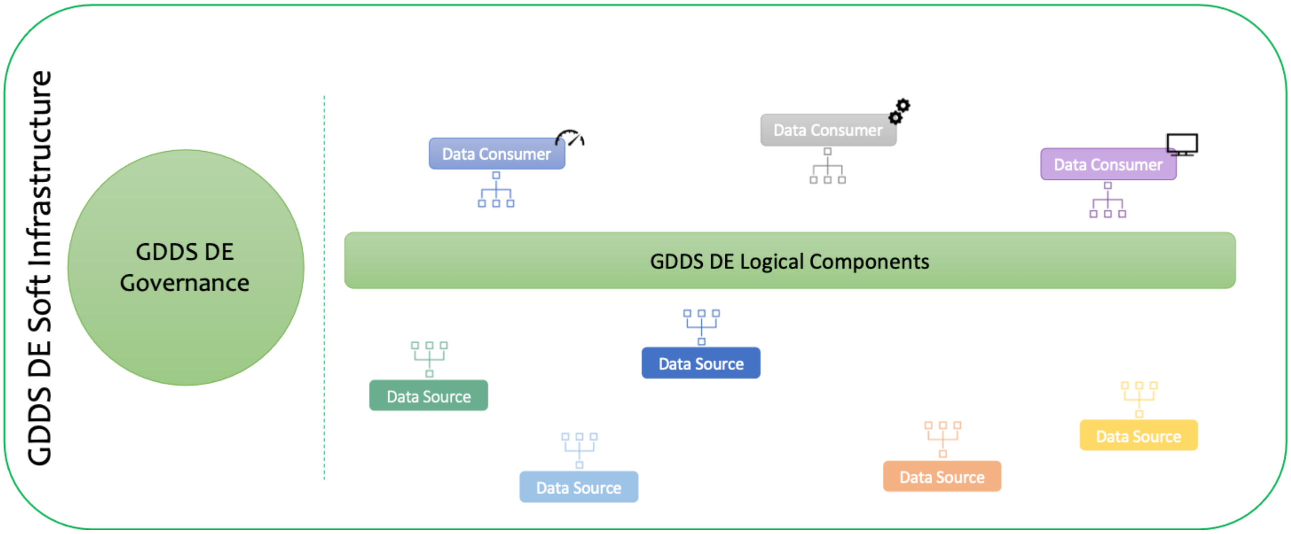


Figure 1 – The GDDS DE Soft Infrastructure

3.5 Governance Challenges

The GDDS DE builds on existing systems, encouraging the development of new elements to fill any gaps, and creating a complex digital environment. The participating systems in the GDDS DE are highly diverse and managed by various organizations, ranging from legacy systems with differing objectives and technological characteristics to systems with diverse content. As a result, the success of the GDDS DE largely relies on proper governance of the ecosystem as a whole. This governance must establish a set of rules and principles to guide the evolution and effectiveness of the ecosystem ensuring to continue delivering the defined ecosystem service, as it navigates through the various changes in the political, social, scientific, and technological environment in which it operates.

While it is out of the scope of this document to detail the specific governance mechanisms which will have to be applied to the GDDS DE (D4.1 describes the GDDS DE governance), it is important to recognize some the main governance challenges which have an impact on the technological framework of the GDDS DE:

a) Who is Part of the Digital Ecosystem: One of the primary concerns for the governance framework is to determine which systems are suitable for inclusion in the GDDS DE. It is important to note that eligibility should focus on the requirements of the organization operating the system, as well as users' needs (i.e., what are the necessary data and resources to address their use cases), rather than technical aspects. In general, it is worth noting that the governance framework must not necessarily impose limitations on eligibility and may instead decide that any system from any organization can be part of the GDDS DE. However, this decision falls within the purview of the governance framework and an appropriate process should be clearly defined.

b) Balancing Belonging vs. Autonomy: The success of a digital ecosystem depends on the ability of participant systems to collaborate and achieve a common value (i.e., the ecosystem service), while also pursuing their own goals. Thus, it is crucial for the governance to establish and manage acceptable behaviors (including the level of openness and transparency), time evolution, and communication and interoperability levels of participating systems. The digital ecosystem must be flexible enough to accommodate different levels of participation and autonomy, which are determined by each system, and guarantee an equal accessibility for all stakeholders. These compromises have a significant impact on the technological solutions supporting the digital ecosystem and should be carefully regulated by the governance framework.

c) GDDS DE Logical Components: As recognized in previous sections, the GDDS DE requires a minimal set of logical components which enable the digital environment where all participating systems can interact. The governance framework should establish a clear process to identify such components; in fact, as part of the adaptation to changes in science/policy/technology contexts, there might be the need to add/dismiss such components. Besides, for each component, the high-level operational governance should be laid out. This includes at least two items: (i) high-level functionalities provided by the component, and (ii) life-cycle process(es) operated by the component.

4 Technical Blueprint of Green Deal Data Space

As described in previous section, the first step for designing the GDDS DE is the definition of its ecosystem service, i.e., its high-level objective. To this aim we utilized the following main inputs:

- The European Strategy for Data.
- The Digital Europe Programme Call.
- The GREAT project proposal.
- The Commission Staff Working Document on Common European Data Spaces.
- The "European Data Spaces - Scientific Insights into Data Sharing and Utilisation at Scale" report.
- Inputs from the consultation with use-cases Task Forces.
- Partners' expertise in major System of Systems (SoS) data sharing initiatives.

The GDDS DE Ecosystem Service is defined as it follows:

Secure, trusted and seamless sharing (i.e., discovery, access and use) of data to support Green Deal applications.

The provision of the ecosystem service is not the only high-level requirement. It concerns the '*what*' dimension of the GDDS DE, but we also need to consider the '*how*' dimension which is related to the recognized high-level values. They can be expressed through a set of general principles acting as requirements and constraints of the ecosystem. In a preliminary phase, we identified the following set of basic principles:

a) *Inclusiveness*: We can expect a high heterogeneity of data systems in terms of supported metadata content and formats, data encoding, coordinate reference systems, ontologies. At least part of this heterogeneity is justified by the specificity of the community that generates and uses those data. Since the driving benefit of a data space is to share *all* the valuable datasets, data systems cannot be excluded only due to their diversity (as long as they do not compromise the overall level of service of the GDDS DE).

b) *Fairness*: we can expect high heterogeneity also in terms of 'species' including big companies, SMEs, public administrations, research and academic organizations, intergovernmental institutions, citizens. A data space should be the common ground where collaboration and competition take place for the benefit of the 'species' but, overall, for the ecosystem to serve data for generating knowledge. Therefore, no privileged access should be granted to anyone at the risk of changing the fairness of the data space.

c) *Autonomy*: we expect that some data sources are already part of other SoS or ecosystems with their own mandate and governance – e.g., European Research Infrastructures, Copernicus Services, Space Agency ground segments, Public Administration systems including INSPIRE. It is necessary to respect such autonomy without imposing, de-iure or de-facto, the exclusive participation in the data space. This is strictly related to the autonomy vs. belonging conflict that will affect any data system. In a Common European Data Space, belonging should be encouraged through soft means mostly based on the overall value of the data space.

In addition to the above basic principles, we identified the following set of architectural design principles for the GDDS technical blueprint:

1. Lower Entry Barrier: the GDDS allows a low entry barrier for both data providers and data consumers.

2. System of Systems: the GDDS is designed as a System of Systems (SoS) to interconnect many independent, autonomous systems, frequently of large dimensions, to satisfy a global goal (i.e., the GDDS DE service) while keeping them autonomous.
3. Standardization and Mediation: the GDDS will rely on interoperability standards, developed at community level, complementing it with mediation/brokering to enable cross-domain interoperability.
4. Data as entry point: the GDDS focuses on the sharing and use of data, independently of how the data is generated (e.g., off-line, on-the-fly, etc.).
5. Loose-coupling: the GDDS DE is enabled by a set of APIs which can be used by data consumers to leverage (and enrich) the GDDS resources.
6. Interoperability/Security Orthogonality: the GDDS security architecture is orthogonal to the GDDS interoperability architecture.

4.1 Orthogonality of data-sharing and security architectures

The general GDDS architecture can be decomposed in a data-sharing architecture describing the structure and interaction of components fulfilling data-sharing requirements, and a security architecture describing the structure and interaction of components fulfilling security requirements. In the GDDS we assume the *orthogonality* of the two architectures, meaning that any change in one of them should not affect the other one. This is a common assumption in software architectures, and it strictly derives from the orthogonality (independence) of data-sharing and security requirements. The advantage of orthogonality is that it allows decomposing architectures handling each aspect separately.

4.2 Architecture Description

A system architecture is the set of “fundamental concepts or properties of an entity in its environment and governing principles for the realization and evolution of this entity and its related life cycle processes” [25]. An architecture is described through an architecture description which is “a set of products that documents an architecture in a way its stakeholders can understand and demonstrates that the architecture has met their concerns” [26].

A complex system cannot be effectively described through a single over-compassing description. It should provide a lot of information ranging from high-level aspects like stakeholders’ interactions with the system, to very low-level aspects such as software object methods, interfaces and technological choices. Different stakeholders would find most of the information unnecessary and too detailed for those aspects they are not specifically interested in. *Viewpoint modelling* addresses this issue providing different views of the same architecture. “*A view is a representation of one or more structural aspects of an architecture that illustrates how the architecture addresses one or more concerns held by one or more of its stakeholders*” [26].

The following paragraphs provide the GDDS DE description according to the following main views adopted in the ISO Reference Model for Open Distributed Processing (RM-ODP) [27]:

- Enterprise Viewpoint
- Information Viewpoint
- Computational Viewpoint
- Engineering Viewpoint
- Technology Viewpoint

4.3 Enterprise Viewpoint

The enterprise viewpoint is concerned with the purpose, scope and policies governing the activities of the specified system within the organization of which it is a part [27]. This viewpoint focuses on the actors, their interactions in scenarios, use-cases and it allows the elicitation of user requirements and then system requirements. As described in previous sections, the design of a Digital Ecosystem is not built around use-cases. However, for the purpose of describing the technical blueprint architecture according to RM-ODP viewpoint modeling it is possible to elicit some high-level functional requirements from the identified GDDS DE Ecosystem Service. The following sections introduce the main actors and high-level functional requirements which were identified.

4.3.1 Actors

It is possible to identify the main actors involved in the creation, enhancement, and growth of the GDDS DE. The identification of the actors (along with their roles and interest in participating) is important to formulate a valid strategy to promote the GDDS DE, triggering the virtuous cycle of utilization/contribution which underpins the success of the ecosystem. The following main actors have been identified:

- Data Provider: the organization which manages one or more Data Sources which are part of the GDDS DE; the interest in participating is mainly the widening of the possible user-base of their resources and gaining more visibility which can help them to achieve their objectives (e.g., funding sustainability, new business opportunities, etc.).

- Intermediate User: the entity (person or organization) which accesses the GDDS DE to use the provided content and generate added-value artifacts (products, services, applications, etc.) which can be exploited by other GDDS DE users, thus enriching the ecosystem itself. Intermediate users benefit from participating mainly by: (i) exploiting the amount of available data which they can use to generate their added-value content, and (ii) offering their content to the rest of the ecosystem users.

- End User: the entity (person or organization) which accesses the GDDS DE to use the provided content. The main benefit for end users stems from the trusted environment guaranteed by the GDDS DE and the possibility to access the GDDS DE through dedicated tools (e.g., desktop/web applications) developed on top of the GDDS DE content.

Figure 2 depicts the identified actors and associates them with a set of basic technical use cases which support the creation, enhancement, and growth of the GDDS DE:

UC1 **Publish Dataset in GDDS DE**: Data Providers publish their datasets in the GDDS DE.

UC2 **Generate GDDS DE-based added-value artifacts**: Intermediate users generate new products, services, applications, etc. utilizing GDDS DE content. This use case might include UC1 in case the newly generated products are published in the GDDS DE.

UC3 **Use of GDDS DE-based Application**: End users utilize dedicated tools (e.g., desktop/web applications) to use the GDDS DE content.

UC4 **Exploit GDDS DE content**: a software agent accesses the GDDS DE to use its content. This generic definition is used to factor out the two common technical use cases of discovering and using GDDS DE datasets.

UC4.1 **Discover GDDS DE datasets**: a software agent discovers GDDS DE datasets provided from different Data Sources.

UC4.2 **Use GDDS DE datasets:** at least the following two use cases can be defined for the use of GDDS DE datasets.

UC4.2.1 **Download GDDS DE datasets:** a software agent downloads datasets from one or more Data Sources in the GDDS DE

UC4.2.2 **Process GDDS DE datasets:** a software agent processes datasets from one or more Data Sources in the GDDS DE

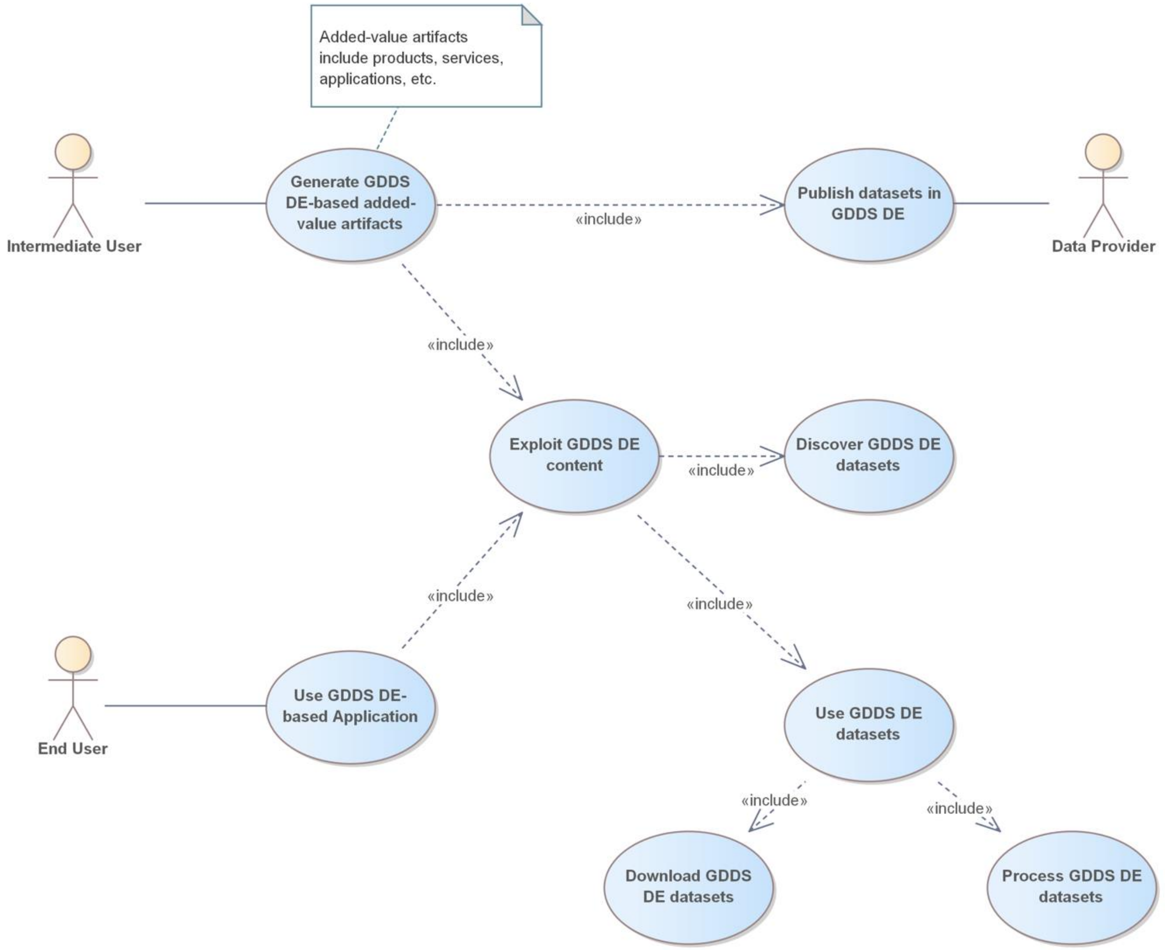

Added-value artifacts include products, services, applications, etc.
Intermediate User
Generate GDDS DE-based added-value artifacts
«include»
Publish datasets in GDDS DE
Data Provider
«include»
Exploit GDDS DE content
«include»
Discover GDDS DE datasets
«include»
«include»
End User
Use GDDS DE-based Application
Use GDDS DE datasets
«include»
«include»
Download GDDS DE datasets
Process GDDS DE datasets

Figure 2 - Main actors involved in the creation, enhancement, and growth of the GDDS DE (UML Use Case diagram)

4.3.2 High-Level Requirements

Table 1 lists the high-level functional and non-functional requirements of the GDDS DE, based on the use cases described in previous section. The list of requirements focuses on interoperability aspects of the GDDS DE.

Table 1- High-level Functional/Non-functional Data Sharing Architecture Requirements

Code	Name	Description
FR1	Data Sources Inventory	The GDDS DE provides an inventory of data sources which populate the digital ecosystem.
FR2	Dataset discovery	The GDDS DE provides discovery of datasets based on different criteria including at least: P 1) geographical coverage expressed as bounding box; P 2) temporal extent expressed as start and end date/hour;

		P 3) keywords present in multiple metadata fields; P 4) data provider expressed as catalog/inventory name;
FR2.1	Dataset discovery protocols (data sources)	The GDDS DE supports different interfaces to discover data from available data sources.
FR2.2	Dataset discovery protocols (clients)	The GDDS DE supports different discovery interfaces to allow clients to discover available data.
FR3	Persistent and Unique Identifiers	Available data in the GDDS DE must be identifiable by a persistent and unique identifier.
FR4	Dataset Access	The GDDS DE provides access to datasets from heterogeneous data sources
FR4.1	Dataset access protocols (data sources)	The GDDS DE supports different interfaces to retrieve data from origin data sources.
FR4.2	Dataset access protocols (clients)	The GDDS DE supports different interfaces which clients can use to retrieve data.
FR5	Dataset Transformation	The GDDS DE provides basic transformation functionalities such as: • sub-setting • interpolation • reprojection on multiple Coordinate Reference Systems • data format transformation Through the GDDS DE, a user can access datasets from different data sources and retrieve them in a common form (same resolution, same CRS, same format, etc.).
FR6	Data processing on Cloud and HPC platforms	The GDDS DE allows client applications to process data on cloud and HPC platforms.
NFR1	Availability	The GDDS DE must ensure the availability of shared data or inform about the temporary unavailability
NFR2	Usability	The GDDS DE be user-friendly for both end users and intermediate users. This includes documentation, user support, training, etc.

4.4 Information Viewpoint

Information viewpoint is concerned with the kinds of information handled by the system and constraints on the use and interpretation of that information [27].

To provide a seamless sharing (i.e., discovery, access, and use) of data to support Green Deal applications, the characteristics of information handled and shared by the GDDS DE is a fundamental aspect. We recognize two main challenges concerning information handled by the GDDS DE:

- Persistent and unique identifiers: available data in the GDDS DE should be identified in a unique and persistent way.
- Heterogeneity: the connected data sources vary largely in terms of service interfaces/APIs, as well as data models and formats of both metadata and data.
- Semantics: the content can be annotated and interpreted according to different semantics (in the form of controlled vocabularies, ontologies, etc.).

4.4.1 Persistent and Unique Identifiers

To be able to track the usage of GDDS DE data, as well as to implement security functionalities (e.g., grant access authorization to data), we must be able to identify available GDDS DE data in a unique and persistent way.

While the concept of unique identifier is self-explanatory, it must be noted that the uniqueness can range from local (i.e., inside a single repository) to global (i.e., an identifier which is unique at the global level). For the objectives of the GDDS DE it is sufficient to have a unique identifier “internal” to the GDDS DE itself (e.g., to trace which data was accessed by whom and when).

The concept of persistent identifiers for documents is not a new one. These were introduced to address the problem of “broken links” – i.e., URLs which become unavailable after some time due, e.g., to a re-organization of a web site structure. In the mid 1990s, several schemes were developed that, rather than relying on the precise address of a document (i.e., the URL), introduced the idea of name spaces for recording the names and locations of documents [28]. Essentially, after registering document identifiers in a central repository, upon an end-user’s request to access a document, the identifier of that document is “resolved” to its exact location (in a transparent way for the end-user) and the document is retrieved.

It must be noted that, conceptually, associating a PUI to data is not as straightforward as to associate it to documents, particularly in the highly heterogeneous context of the GDDS DE. We can recognize at least two main challenges which differentiate this association from the one with documents:

1. Data hierarchy and granularity can be very different for different data sources depending on both the type of data which is shared and the scientific domain.

2. Data access services/APIs often allow the access to: (i) a (temporal and/or spatial) subset of the entire data, (ii) a different encoding format of the data, (iii) some simple transformation of the data (e.g., change of CRS), etc.

Essentially, in the context of the GDDS DE it is not possible to associate a PUI to a single file in the same way this is done with documents. To address this, we introduce the concept of Logical Resource, which represents an abstract element which is used to identify the data which is shared[4]. Figure 3 depicts the UML class diagram of the Logical Resource and its associated elements. The PUI is associated (1 to 1) with the Logical Resource, which in turn is described by a set of Metadata (among which, the PUI itself). Finally, the Logical Resource can be instantiated by two concrete elements (Dataset Collection, and Dataset) that represent the concrete artifacts shared by a Data Source (see Figure 5).

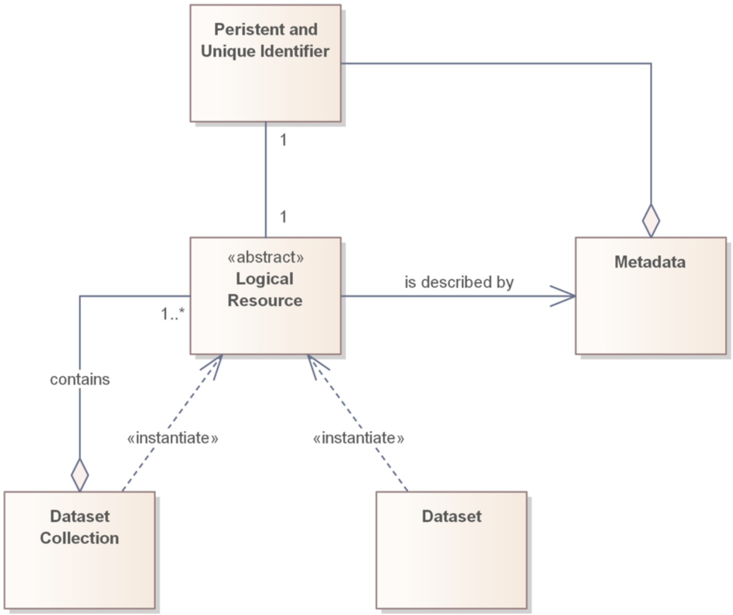


Figure 3 - Persistent and Unique Identifier of GDDS DE Logical Resource

4.4.2 Heterogeneity

The GDDS DE aims to facilitate the sharing and use of geospatial data related to Green Deal, a domain which is characterized by a high level of heterogeneity, with many already existing data sharing initiatives that offer their resources to diverse consumers, which mirrors the current state of (geospatial) data sharing globally. Although a certain level of

[4] It is worth to note that this definition is aligned with the relationship between URI and Resources in the Web, as described in https://www.w3.org/TR/webarch/#id-resources

standardization can be expected, based on a well-defined governance process for the identification of relevant standards to be supported, the GDDS DE must take care of all the mediation, harmonization and transformation actions needed to make geospatial data easily discoverable, accessible, and usable.

This means that the GDDS DE must be able to handle different service interfaces and metadata/data models for discovery and access. Based on the experience of other large multidisciplinary data sharing initiatives (e.g., the Global Earth Observation System of Systems, GEOSS) it is possible to list some of the relevant service interfaces and metadata/data models for discovery and access (see Annex A). The GDDS DE must be able to connect to data sources which utilize such interfaces and data/metadata models for data sharing. Besides, the GDDS DE must expose such interfaces and data/metadata models towards applications which want to access GDDS DE resources. For example, the GDDS DE must enable a visualization application requesting data according to the OGC Web Map Service (WMS) interface to retrieve data provided by a data source that shares its resources via the OGC Web Coverage Service (WCS) interface or another service interface (e.g., THREDDS Server).

4.4.3 Semantics

Many of existing data sharing systems and initiatives make use of semantic artifacts for the description of their shared data, including the use of controlled vocabularies, ontologies, etc. The use of such artifacts addresses the need to support a higher level of interoperability, i.e., semantic interoperability. This aims to ensure that the meaning of exchanged data and information are preserved and understood throughout exchanges between parties, in other words “what is sent is what is understood”.

There exist several initiatives which develop semantic artifacts and services, both at a domain level (e.g., WHOS Hydrological Ontology[5], AGROVOC[6], SeaDataNet Vocabularies[7], NetCDF CF[8], etc.) and at a general-purpose level (GEMET[9], EuroVoc[10], GCMD Keywords[11]). Sometimes, even data sharing initiatives from the same domain utilize different semantic artifacts to describe their data.

Again, the heterogeneity characterizing the Green Deal domain plays a crucial role. In fact, on one side, it is important to use descriptions based on semantics to address Green Deal variety and differences. These descriptions should be preserved when data is shared in the GDDS DE and made available to Data Consumers, giving them all necessary information to assess if the available data meets their needs.

[5] https://community.wmo.int/en/whos-hydrological-ontology

[6] https://www.fao.org/agrovoc/

[7] https://vocab.seadatanet.org/search

[8] http://cfconventions.org/cf-conventions/cf-conventions.html

[9] https://www.eionet.europa.eu/gemet/en/about/

[10] https://eur-lex.europa.eu/browse/eurovoc.html?locale=en

[11] https://www.earthdata.nasa.gov/learn/find-data/idn/gcmd-keywords

On the other hand, the use of different semantic artifacts and services gives origin to the need of aligning and mapping the contents of such a heterogeneous environment. This task (aligning and mapping) is very demanding, mainly due to the conceptual aspect of it. In fact, aligning and mapping semantic concepts in a cross-domain environment requires a very deep scientific knowledge of the different domains. Even more challenging is the development of tools which can automatically perform such a task. This still represents a research topic with no consolidated results yet available. There are promising techniques (e.g., using AI/ML-based solutions) that might soon provide advances in this field and could be accommodated in this design as a new Facilitator Component (see 4.5.2).

4.5 Computational Viewpoint

Computational viewpoint is concerned with the functional decomposition of the system into a set of objects that interact at interfaces - enabling system distribution [27]. In the case of this technical blueprint architecture, this viewpoint describes the set of logical components which enable the GDDS DE. As depicted in Figure 4, such components can be classified in two main categories: Core and Facilitators. The former identifies the logical components which are critical for the existence of the DE; the latter category identifies the components which facilitate the use of data available in the DE. Both Core and Facilitators components expose Web APIs which data consumer tools can use to exploit the GDDS DE data.

The distinction between Core and Facilitators components is important considering the evolutionary nature of Digital Ecosystems. The Core components are expected not to evolve at a rapid pace, they constitute the foundation of the GDDS DE and are expected to be relatively stable in terms of basic functionalities. On the other hand, Facilitators are designed to enable an as seamless as possible use of the GDDS content. These components are expected to evolve (both in number and in functionalities) more rapidly in response to both users' needs and the emergence of new technologies. In fact, as explained in previous sections, the GDDS DE technical blueprint must be able to cope with a rapidly changing technological environment where we expect the emergence of new technologies, enabling now unpredictable scenarios. In such a context, it is crucial to be able to rapidly adapt and incorporate such changes. Besides, Facilitators can be added incrementally, allowing a smooth growth of the GDDS DE.

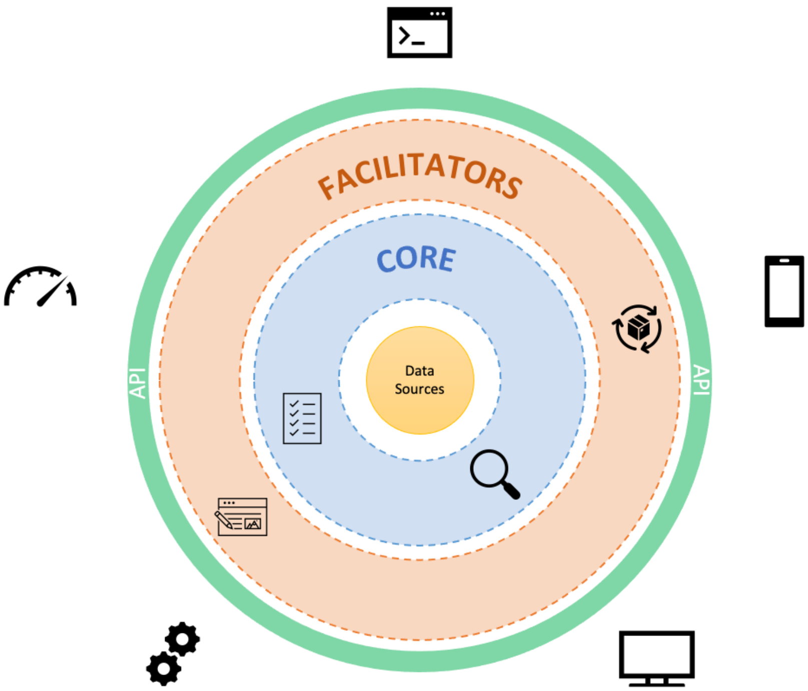


Figure 4 - Core and Facilitators

The initial entity to be considered for modelling the Core and Facilitators Logical Components is the Data Source. This represents a Web-based system which shares its data in the GDDS DE. For the scope of this document, it is useful to model the Data Source (Figure 5) as a component which exposes two interfaces: (i) Dataset Discovery, and (ii) Dataset Access. A Data Source is managed by a Data Provider.

Interface
In the remainder of this section, interface is generically utilized to express the set of operations which a system/component exposes as well as the data models and formats utilized for message (request/response) exchange.

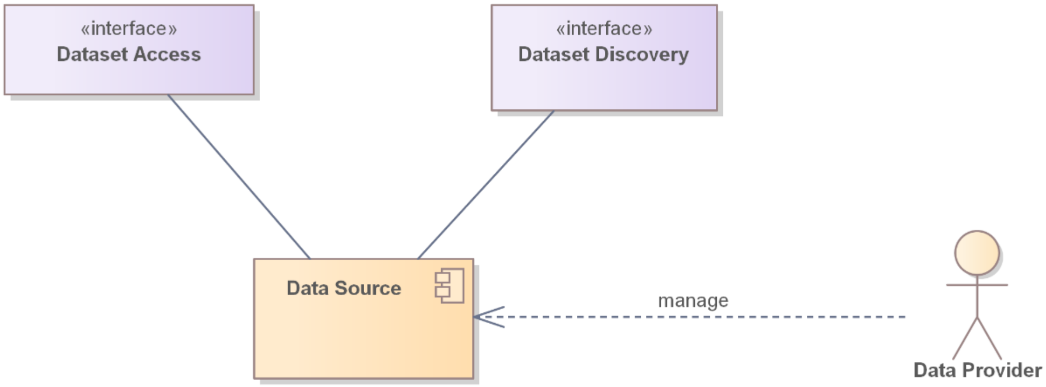


Figure 5 - Modelling of Data Source Component

Another key entity to model is the Data Consumer. As depicted in Figure 6, a generic Data Consumer component is modeled as a component which utilizes GDDS DE interfaces, i.e., the set of interfaces exposed by the GDDS DE. A generic Data Consumer component can be further specialized to highlight which Actors are associated with different type of Data Consumer components. The first level of specialization differentiates between GDDS DE Logical Component and Third-Party Component. This second type is associated with Intermediate Users that can develop different types of components accessing and exploiting the GDDS DE content. In particular, the Third-Party Component can be specialized in Client Application and Middleware. The former type of component refers to all tools which target the End Users, that use them to access and exploit the GDDS DE content. The latter type is instead used to describe those components which are not directly used by End Users but provide added-value services (built on top on the GDDS DE content) which can be exploited by other Third-Party Components. It is worth to note that this is key aspect for the growth of GDDS DE.

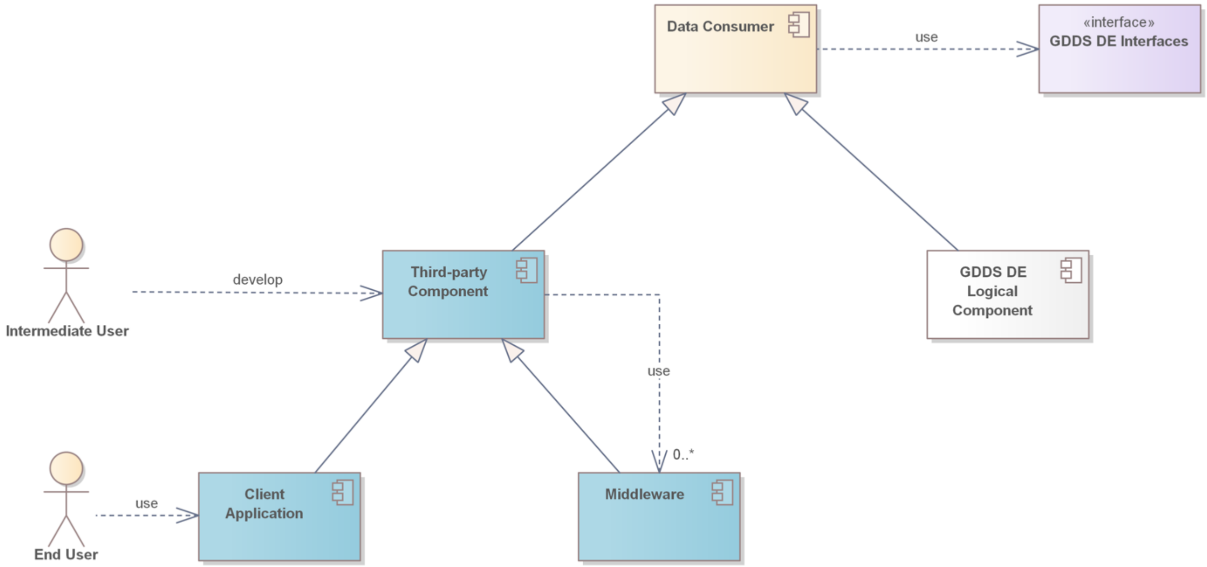


Figure 6 - Modelling of Data Consumer Component

The following sub-sections describe the initial set of logical components which were identified for the Core and Facilitators categories.

4.5.1 Core Logical Components

For the GDDS DE to exist, it is first necessary to know which data are available. To this aim, the first and foremost component which is required is a Registry of Data Sources. This component is in charge of collecting the list of Data Sources which are part of the GDDS DE, along with all necessary interoperability interfaces exposed by each Data Source. The Registry of Data Sources must expose two interfaces: (i) a Data Source Register interface which is used to register Data Sources, and (ii) a Data Source Inventory interface which allows the retrieval of registered Data Sources and the associated interoperability information.

The second component to enable the discovery of available data is a Data Catalogue. This component is in charge of providing a set of Uniform Discovery interfaces. Such interfaces can be used by data consumers to discover available data in the GDDS DE. To do so, this component connects to the different Data Sources listed by the Registry of Data Sources and utilizes the Data Discovery interface which each Data Source exposes to retrieve metadata from that Data Source. All necessary mediation and harmonization functionalities are implemented by the Data Catalog. The second interface exposed by the Data Catalog is a Metadata Update interface; this allows to modify/enrich original metadata with additional information. This interface is utilized by the Status Checker component. The task of this component is to check the status of availability of the different Data Sources and update the correspondent metadata with this information. This allows, on one hand, data consumers to know if the discovered data is available and, on the other hand, to inform data providers about possible issues related to the access of their data.

These first three Core Logical Components (Data Source Registry, Data Catalog and Status Checker) address the very basic requirements for the GDDS DE, i.e., what data is available in the GDDS DE. Data access requires data consumers to utilize the different Data Access interfaces exposed by the different Data Sources.

As described in the information viewpoint section (see 4.4.1), data in the GDDS DE must be identified by persistent and unique identifiers (PUIs). To this aim, we introduce two logical components: the PUI Provider, and the PUI Resolver. The former is tasked with providing a PUI for Logical Resources in the GDDS DE, while the latter resolves a PUI to return the corresponding Logical Resource representation.

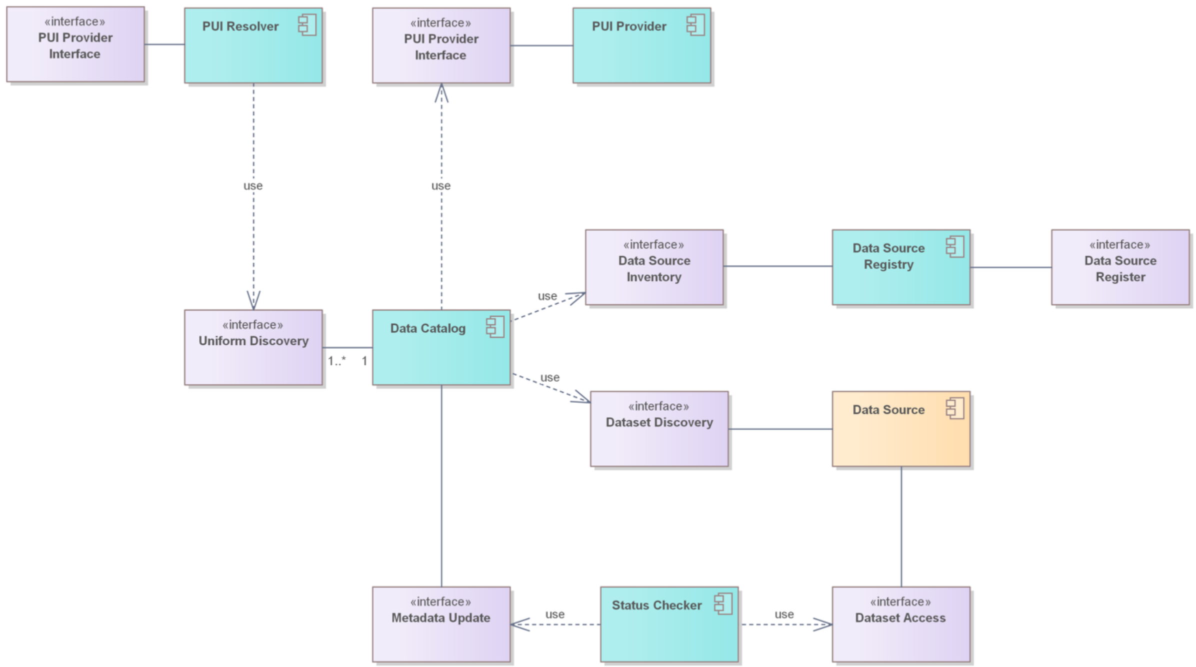


Figure 7 – UML Diagram of Initial Set of Core Logical Components of the GDDS DE

Figure 7 depicts the Core Logical Components, and their interactions at the interface level (with each other and with the Data Sources), while Table 2 lists these components with a brief description and the high-level requirements they address.

Table 2 - List of Initial Set of Core Logical Components

Component	Description	Requirements
Registry of Data Sources	Allows the registration and retrieval of GDDS DE Data Sources	FR1
Data Catalog	Allows the discovery of available data from registered Data Sources.	FR2
Status Checker	Checks the availability status of GDDS DE Data Sources	NFR1
PUI Provider	Provides a PUI for Logical Resources of the GDDS DE	FR3
PUI Resolver	Resolves a PUI for Logical Resources of the GDDS DE	FR3

4.5.2 Facilitators Logical Components

The aim of the logical components in this category is to facilitate the use of the GDDS DE content. With respect to Core Logical Components, Facilitators are expected to evolve (both in number and in functionalities) more rapidly, in response to both users' needs and the emergence of new technologies. Figure 8 depicts the component diagram of the initial set of Facilitators Logical Components of GDDS DE and their main interactions.

At this stage of the design, it is already possible to identify a first set of Facilitators addressing the main obstacles for use of data in the heterogeneous context of the GDDS. One of the main entry barriers to the use of data is represented by data access, i.e., the possibility for data consumers to obtain the required data in a form which the data consumer can use. Two main issues must be addressed for facilitating data access:

1. Data Access interface heterogeneity: the different Data Sources use different Data Access interfaces; therefore, data consumers must implement these interfaces to be able to access the data.
2. Data Form heterogeneity: this includes at least data format encoding, coordinate reference system (CRS), spatial and temporal extent. Data consumers need not only to access (download) the data, but they need it in a form which is suitable for their needs. Again, Data Sources provide, through their Data Access interfaces, a subset of all possible data forms required by the different data consumers which must implement all necessary transformations before using the data.

To address these issues, we introduce the Dataset Transformer component. This provides all mediation, harmonization and transformation functionalities which are needed to shift the burden of dealing with the above-described issues from data consumers. The Dataset Transformer provides a set of Uniform Data Access interfaces; each of these interfaces will comply with one standard recognized by the GDDS DE. Data consumers can utilize the preferred Uniform Data Access interface to request data access according to their needs (data format encoding, CRS, etc.). The Dataset Transformer retrieves the requested data from the origin Data Source, utilizing the Dataset Access interface exposed by the Data Source, and (if needed) executes the necessary transformations to comply with data consumer’s request (e.g., data format encoding transformation, CRS transformation, etc.).

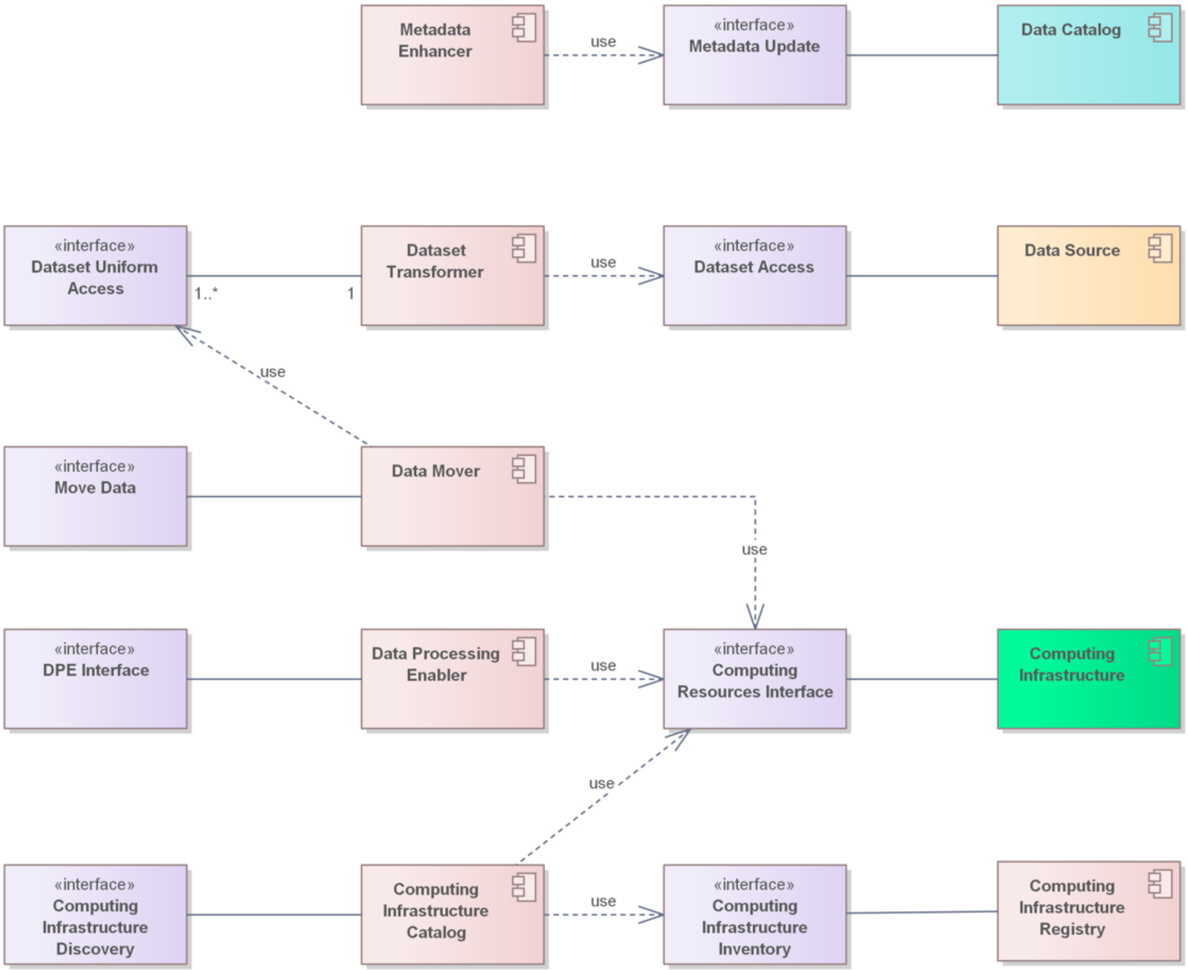

Metadata Enhancer
use
«interface» Metadata Update
Data Catalog
«interface» Dataset Uniform Access
1..*
1
Dataset Transformer
use
«interface» Dataset Access
Data Source
use
«interface» Move Data
Data Mover
use
«interface» DPE Interface
Data Processing Enabler
use
«interface» Computing Resources Interface
Computing Infrastructure
use
«interface» Computing Infrastructure Discovery
Computing Infrastructure Catalog
use
«interface» Computing Infrastructure Inventory
Computing Infrastructure Registry

Figure 8 - UML Diagram of Initial Set of Facilitators Logical Components of GDDS DE

The Dataset Transformer logical component facilitates data access, supporting the traditional pattern of discovery and download of data, which is then locally processed to generate an added-value product. However, in the context of Big Data which characterizes the current landscape of the data economy, this pattern covers only partially the users' needs. In fact, it is often very inefficient (and sometimes impossible) to download all the required data for an application. Besides, the computational, storage and network requirements for handling a Big Data-based application are very hard to be met by a local data center, in terms of both infrastructure management and cost. Cloud and HPC platforms offer the necessary capabilities to cope with Big Data requirements. To be able to use such platforms for data processing, data consumers must have access to the requested data on such platforms. To this aim we introduce the Data Mover facilitator. This component must take care of implementing all required actions to make the requested data available (in the desired form) on the requested platform. A specific interface is exposed by the Data Mover for requesting the data, and in turn the Data Mover will utilize the Dataset Transformer interface for retrieving the requested data and move it to the requested Cloud/HPC platform. Such platforms expose their functionalities through a set of interfaces which are usually broadly characterized, according to the type of resources they to manage, as:

- IaaS (Infrastructure as a Service): the capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources [29].
- PaaS (Platform as a Service): the capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider [29].
- SaaS (Software ad a Service): the capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface [29].

However, at this stage of the design it is sufficient to model a Cloud/HPC platform as a Computational Infrastructure component which exposes a generic Computing Resources Interface encompassing IaaS, PaaS, and SaaS capabilities. Each supported Computing Infrastructure, along with the specification of its Computing Resources Interface, is registered in a Computing Infrastructure Registry and can be discovered via a Computing Infrastructure Catalog.

The Facilitators introduced so far simplify data retrieval (either to local facilities or to Cloud/HPC platforms) for data consumers. However, to process such data in Cloud/HPC platforms, data consumers still need to interact with the different Computing Resources Interfaces to set up an execution environment which satisfies their needs (e.g., instantiate virtual machines, set up the execution framework with the required libraries, etc.). Most of these tasks can be automated and implemented against the different Computing Resources Interfaces exposed by the Computing Infrastructures in the GDDS DE. A new Facilitator is introduced to this aim: the Data Processing Enabler. Such a component exposes an interface which can be invoked by data consumers to submit the execution request of their own specific algorithm implementations, along with some basic information about the requirements (e.g., required CPU and memory, execution framework description, etc.) and the input data to process. The Data Processing Enabler takes care of setting up the execution environment on the Cloud/HPC platform, triggers the execution and saves the output.

Finally, we introduce another facilitator: the Metadata Enhancer. This component is in charge of enhancing the metadata which are available via the Data Catalog. Such a component will in general enhance the usability of the GDDS DE by the different Data Consumers. In fact, such enhanced metadata can be exploited to facilitate the discovery of required datasets by the different Data Consumers. Metadata might be enriched with, e.g., fit-for-purpose information or with other relevant Data Consumer-driven information.

Table 3 summarizes the facilitators introduced and the main requirements they address.

Table 3 - List of Initial Set of Facilitators Logical Components

Component	Description	Requirements
Dataset Transformer	Allows data access from different Data Sources according to a common Data Form (data format, CRS, etc.)	FR4 FR5
Computing Infrastructure Registry	Allows the registration of GDDS DE supported Computing Infrastructures.	FR6
Computing Infrastructure Catalog	Allows the discovery of GDDS DE supported Computing Infrastructures.	FR6
Data Mover	Makes available dataset from different Data Sources according to a common Data Form on the supported Computing Infrastructures.	FR6
Data Processing Enabler	Allows the execution of a Data Consumer's algorithm on different Computing Infrastructures.	FR6
Metadata Enhancer	Allows to enrich metadata in the Data Catalog.	FR2, NFR2

4.6 Engineering Viewpoint

Engineering viewpoint is concerned with the infrastructure required to support system distribution [27]. For the purpose of this document, it is useful to introduce the following types of Nodes:

- Computing and Storage Node: this represents a traditional Data Center node, which can be used by the node owner to store data and deploy one or more data services. This type of node is represented in gray in Figures 10-11.

- Cloud/HPC Node: this type of node represents Cloud and HPC platforms and is assumed to always provide a Computing Infrastructure component (as well as its associated Computing Resources Interface providing IaaS/PaaS/SaaS capabilities). The main difference of this type of node with respect to Computing and Storage Node is the availability of the Computing Resources Interface. Thus, differently from the Computing and Storage Node, external applications/developers can exploit the available IaaS/PaaS/SaaS capabilities to store data and deploy one or more data services on this type of nodes. This type of node is represented in green in Figures 10-11.

- An End User Device Node: this represents a node hosting End User's Client Applications. It can be a desktop, or a mobile device. This is characterized by a very limited amount of computational, storage and network bandwidth resources. This type of node is represented in light blue in Figures 10-11.

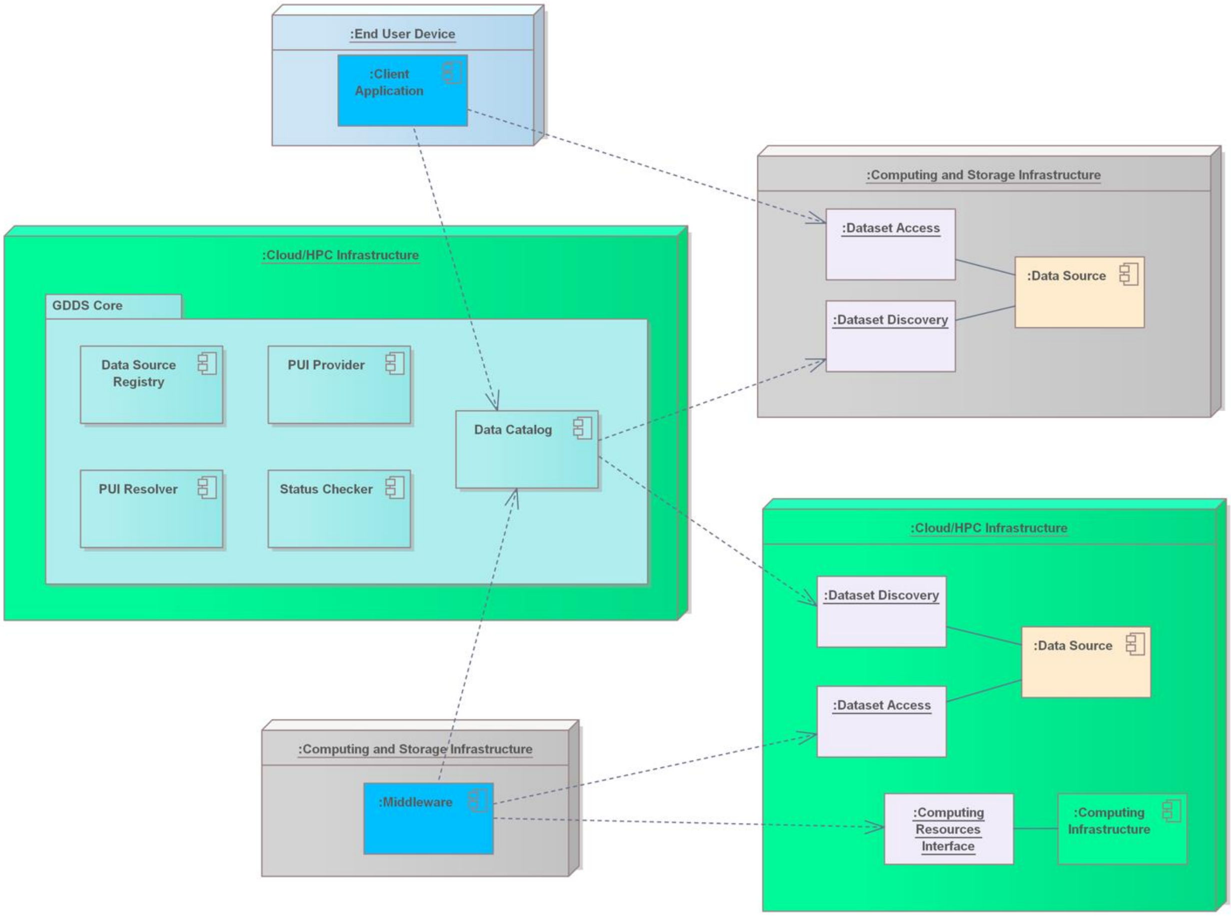


Figure 9 - Example of Core Components Engineering Diagram

Figure 9 depicts a possible deployment scheme for the Core Components of the GDDS DE. These are centralized components and are deployed on a Cloud/HPC Infrastructure node. Besides, the diagram depicts a possible scenario with a couple of Data Consumers (a Client Application and a Middleware) and Data Sources. The Client Application is deployed on an End User Device node, while the Middleware is deployed on a Computing and Storage Infrastructure node. One of the two Data Sources is deployed on a Computing and Storage Infrastructure node while the other one on a Cloud/HPC Infrastructure node. In this simple example, the Client Application must discover the data of interest and retrieve it for displaying to the End User. In addition to the discovery and retrieval of data, the Middleware must execute some additional processing, utilizing a Cloud/HPC Infrastructure to take advantage of its scalability.

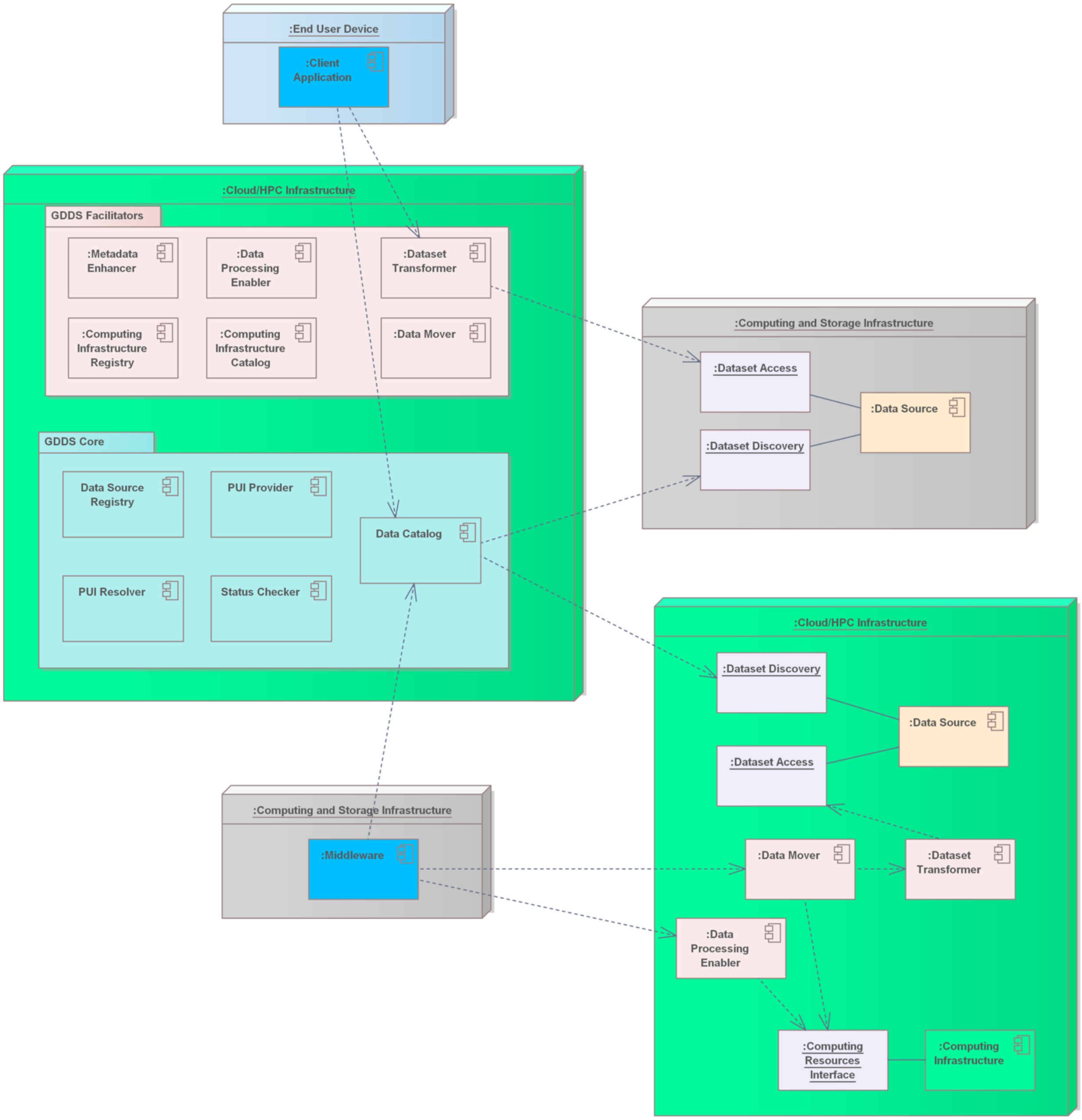


Figure 10 - Example of Facilitators Components Engineering Diagram

Figure 10 extends Figure 9 with the deployment of the Facilitators Components, depicted in red. Some of the Facilitators Components are centralized (Metadata Enhancer, Computing Infrastructure Registry and Computing Infrastructure Catalog), while others (Dataset Transformer, Data Mover, Data Processing Enabler) benefit from a distributed deployment approach. In fact, these latter components are specifically targeted to work on (big) data available in the GDDS DE and should be deployed as close as possible to the data. In Figure 10 they are deployed both on the central node and on the Cloud/HPC Infrastructure node (where a Data Source is deployed as well).

In both diagrams, the discovery phase for both the Client Application and the Middleware is the same, going through the Data Catalog of the GDDS DE which is connected to the Dataset Discovery interfaces of the Data Sources. Instead, the access and use phase is different. Without the Facilitators Components, the Client Application must directly interconnect with the Data Source for the retrieval of the required data. This step might be difficult or inefficient because the Client Application must perform all tasks which the Data Source might not be able to accomplish (e.g., format transformation, CRS conversion, etc.) and these tasks are executed on an End User Device node (therefore with limited amount of computational, storage and network bandwidth resources). As depicted in Figure 10, the introduction of the Dataset Transformer allows the Client Application to retrieve the required data through this component, which runs on a Cloud/HPC Infrastructure node (therefore taking advantage of large computational, storage and network bandwidth resources), implements all necessary tasks to transform the data as requested by the Client Application and finally returns the data to the Client Application itself, alleviating it from the transformation tasks execution and, in most cases, reducing the amount of data to be downloaded.

The Middleware takes advantage of the distributed deployment of the Facilitators Components. The Dataset Transformer can execute its tasks directly on the Cloud/HPC Infrastructure node where the discovered data is already available, the Data Mover stores the transformed data on the same node and finally the Middleware requests the execution of its processing algorithm, through the Data Processing Enabler, on the data previously stored. This way, in this example, all data processing tasks (transformation and specific algorithm execution) are carried out without the need to move data.

4.7 Technology Viewpoint

Technology viewpoint is concerned with the choice of technology to support system distribution [27].

Since the implementation details are out of the scope of this document, we provide in this section a short and non-comprehensive list of possible technological solutions which could be used/extended/combined to implement the logical components we described. The aim is not to suggest the use of the listed technologies, but to show the technical feasibility of the proposed logical components.

The functionalities offered by the Data Catalog, Dataset Transformer, Data Sources Registry and Status Checker logical components are provided in the context of the Global Earth Observation System of Systems (GEOSS) by the components of the GEOSS Platform (former GEOSS Common Infrastructure, GCI). The GEO Discovery and Access Broker (GEO DAB) [18] implements a brokering framework for data discovery and access. The GEO DAB implements the necessary mediation, harmonization, and distribution functionalities to allow data providers to share resources without having to make major changes to their technology or standards. Presently GEOSS Platform, through the GEO DAB, brokers more than 180 autonomous data sources. Based on the same brokering technology, the WMO Hydrology Observing System (WHOS) [30] implements a brokering framework for linking hydrologic data providers and users through a hydrologic information system of systems enabling data registration, discovery and access. The GEOSS Yellow Pages service implements the simplified registration process for new Data Providers. The GEOSS Service Status Checker is the component, developed by USGS/FGDC, which implements an automatic mechanism to monitor, diagnose and alert data providers and users on the Health status of the web services provided by the GEOSS Platform.

Other brokering technologies were developed in other contexts. PANGAEA has set up a brokering framework applicable to earth and environmental sciences [31]. The framework is used since 2007 for the ICSU World Data System (WDS) data portal. EUDAT has elaborated a number of infrastructural tools among them a metadata discovery service - B2Find [32] - which is used to harvest metadata from research data collections from EUDAT data centers and other repositories. The Climate Data Store[12] (CDS) of the Copernicus Climate Change Service (C3S) implements a broker component to schedule and forward data and compute requests to the appropriate data repository (or the compute layer) via a set of adaptors, translate data and computation requests issued by the broker on behalf of the user into requests that are understood by the infrastructure of each of the data providers.

As far as Persistent and Unique Identifiers, one of the most widely used implementation is the DOI[13] (Digital Object Identifier). A large scale system implementing PUIs functionalities was developed to publish and distribute the extensive archive of climate model output generated by the Coupled Model Intercomparison Project Phase 6 (CMIP6) [33].

Data Mover can be based on the many technologies which enable cloud-native distributed storage, e.g., Longhorn[14], IOMesh[15], Ceph[16], etc. Several technologies were developed in the last years to simplify the use of multiple cloud platforms. Terraform[17] is a tool to manage the entire lifecycle of infrastructure using infrastructure as code on multiple cloud providers. Kubernetes[18] is an open-source container-as-a-service (CaaS) framework to automate application deployment, scaling, and operations. Now part of the Cloud Native Computing Foundation, Kubernetes enables application developers to leverage capabilities like self-monitoring, process automation, container balancing, storage orchestration, and more. These and/or other technologies can be combined to develop the Data Processing Enabler component; one example of such a combination to support

[12] https://www.ecmwf.int/en/newsletter/151/meteorology/climate-service-develops-user-friendly-data-store

[13] https://www.doi.org/the-identifier/what-is-a-doi/

[14] https://www.rancher.com/products/longhorn

[15] https://www.iomesh.com/

[16] https://ceph.io/en/

[17] https://www.terraform.io/

[18] https://kubernetes.io/

the execution of scientific environmental models in a multi-cloud environment is the Virtual Earth Laboratory (VLab) [34] [35].

4.8 Security and Trust Architecture

Before describing the blueprint for GDDS DE security and trust architecture, it is useful to introduce a classification of security services, as provided by [36], as well as their definitions as in [37]:

- Authentication: The process of verifying a claim that a system entity or system resource has a certain attribute value. […] Security services frequently depend on authentication of the identity of users, but authentication may involve any type of attribute that is recognized by a system.
- Access Control: Protection of system resources against unauthorized access; a process by which use of system resources is regulated according to a security policy and is permitted only by authorized entities (users, programs, processes, or other systems) according to that policy.
- Confidentiality: The property that data is not disclosed to system entities unless they have been authorized to know the data.
- Integrity: The property that data has not been changed, destroyed, or lost in an unauthorized or accidental manner.
- Non-repudiation: protection against false denial of involvement in an association (especially a communication association that transfers data).

Of course, in general, the type of services which are really necessary and how to implement them depends on the system to be realized. Therefore, the realization of a security system typically follows a stepwise approach, from the analysis of the system to the implementation of the identified security measures. A well-adopted methodology is described in [38]:

a) Identify what you are trying to protect.

b) Determine what you are trying to protect it from.

c) Determine how likely the threats are.

d) Implement measures which will protect your assets in a cost-effective manner.

e) Review the process continuously and make improvements each time a weakness is found.

It is worth noticing here the emphasis on the cost-effectiveness of the implemented security measures. That is, the cost[19] of protecting the system against a threat should be

[19] Cost in this context should be remembered to include losses expressed in real currency, reputation, trustworthiness, and other less obvious measures. [38]

less than the cost of recovering if the threat were to strike the system [38]. Another important aspect to consider is that security measures generally concern both the technological and the organizational (governance) domains.

4.8.1 Access Control of Digital Content

In general, two broad categories of Access Control approaches can be identified:

- Digital Rights Management (DRM): this approach refers to a complete management of digital rights both for access to the digital content and for the management of the digital content itself once accessed and transferred locally. Therefore, DRM is an end-to-end solution, protecting the digital content during its entire lifecycle.

- Remote Access Control: this approach refers to the protection of digital content only during the access phase. Systems which realize this approach protect the digital content until this is transferred to the consumer.

The Remote Access Control approach is therefore more limited than DRM. However, it is also less impacting on participants in the system. In fact, typically, DRM-based systems require the use of specific technologies to provide and use the digital content. On the other hand, Remote Access Control-based systems do not constraint participants to the use of specific technologies. To exemplify this, consider the transfer an e-book content. With DRM, the transferred e-book will be usable (readable) only by applications which support the utilized DRM technology. With Remote Access Control, once transferred, the e-book is readable by any e-book reader application.

However, it is worth to note that, first, the two approaches are not mutually exclusive (in fact, a Remote Access Control can be part of a wider DRM system). Besides, it must be noted that the two systems differ not in the type of security they offer, but in how this security is realized. In fact, the DRM realizes it via a technology which enforces the respect of the content usage license after transferring the content. In the Remote Access Control approach, the license still applies to the transferred content, but its respect is left with the user.

4.8.2 Security and Trust in GDDS DE

At this stage of the design, we focus on the computational aspect (viewpoint) of the security and trust architecture of the GDDS DE.

We can identify the following Actors in the security and trust architecture:

- Data Owner: the entity (person or organization) which owns resources in the GDDS and can grant access and usage rights for those resources.

- Data User: the entity (person or organization) which accesses the GDDS DE content.

- Trusted Middleware: a component which takes part in resource access operation (e.g., a data transformer). These components are not assigners or assignees of policies, they are trusted; typically, GDDS DE Core and Facilitator components are trusted middleware.

Table 4 - High-level Functional/Non-functional Security Architecture Requirements

Code	Name	Description
SFR1	Authentication	Data Users can authenticate in the GDDS DE
SFR1.1	Sigle-Sign-On (SSO)	It is possible to authenticate once to access (if authorized) all resources in the GDDS DE.
SFR1.2	Multiple Identity Providers	The GDDS DE supports authentication via multiple Identity Providers (with different levels of trust)
SFR1.3	Identity Management	Encompasses the entire lifecycle of user account management (creation, modification, suspension, etc.).
SFR2	Authorization	The GDDS DE resources are protected and the access to such resources is restricted to authorized Data Users.
SFR3	Policy Management	Data Owners can create/modify/delete policy associated to the data they share in the GDDS DE.
SFR4	Integrity	The GDDS DE verifies that exchanged data has not been altered.
SFR5	Non-repudiation	The GDDS DE protects against false denial of data exchanges.

At the computational level, the main security and trust architectural choices are the following:

- Decoupling of Authentication and Authorization: the business logics for authentication and for authorization are separated. This is a good practice in general, but even more in a distributed system like GDDS DE where the authorization policies are defined locally (by the different Data Owners).

- Authorization Framework: the authorization (i.e., access control) framework is based on the Remote Access Control approach and compliant with the XACML framework. This choice is driven by the recognition that such an approach has a minor impact on GDDS DE participants, allowing in the initial phase easier on-boarding. As noted in 4.8.1, the Remote Access Control approach can be seen as part of a wider end-to-end DRM which can be introduced at a later stage.
- Logical Resource: the GDDS DE Logical Resource (introduced in 4.4.1) represents an abstract element which is used to identify the data which is shared; therefore, the GDDS DE Logical Resources are the entities which must protected. The GDDS DE Logical Resource is the intersection between the orthogonal security and data-sharing architectures.

Figure 11 depicts the main logical entities of the security and trust architecture, based on the well-known and widely adopted XACML framework [39].

A Data User that wants to execute an action on a resource requests access using an application (Requester) must go through the security gate of the Gatekeeper. This acts as the Policy Enforcement Point (PEP) of the XACML framework and oversees all necessary operations to filter access requests based on the Data Owner's defined policies.

The response generated by the PEP (deny/permit) requires a decision process which is, partially, specific for the required resource and action and, partially, general (e.g., the application of formal rules). It is therefore useful to separate the components which implement the specific and the general business logics; this way, for each resource, only the specific business logic must be provided whereas the general business logic part is provided by a common logical component for all Gatekeepers. The Context Handler is part of the Gatekeeper and implements the specific business logic, e.g., extracting from the access request the necessary information – about the requested action, resource, Data User's identity, etc. – and expressing it according to the Authorizer language. The Authorizer acts as the Policy Decision Point (PDP), implementing the generic business logic part of the decision process. It evaluates an access request based on the policies, provided by the Policy Provider (Policy Administration Point – PAP). The Authorizer response includes the result of the decision process (permit/deny) and a set of obligations which must be satisfied to fulfil the policy associated with the resource.

Upon receiving the Authorizer (PDP) response, the Gatekeeper (PEP) checks the fulfillment of the obligations. Typical examples of obligations are actions to be carried out such as the use of specific security services (integrity, confidentiality, etc.). An Obligation Provider is a component which is able to: (i) verify if a requested obligation is supported, and (ii) implement the obligation.

The necessary attributes to pass the access control of the Gatekeeper are provided by specific Attribute Providers. The Federated Attribute Provider is tasked with mapping attributes from the different Attribute Providers to a common representation in GDDS DE. Among possible Attribute Providers, the Identity Provider provides the Data User's authentication proof and the related identity attributes. The Identity Provider acts as the Policy Information Point (PIP).

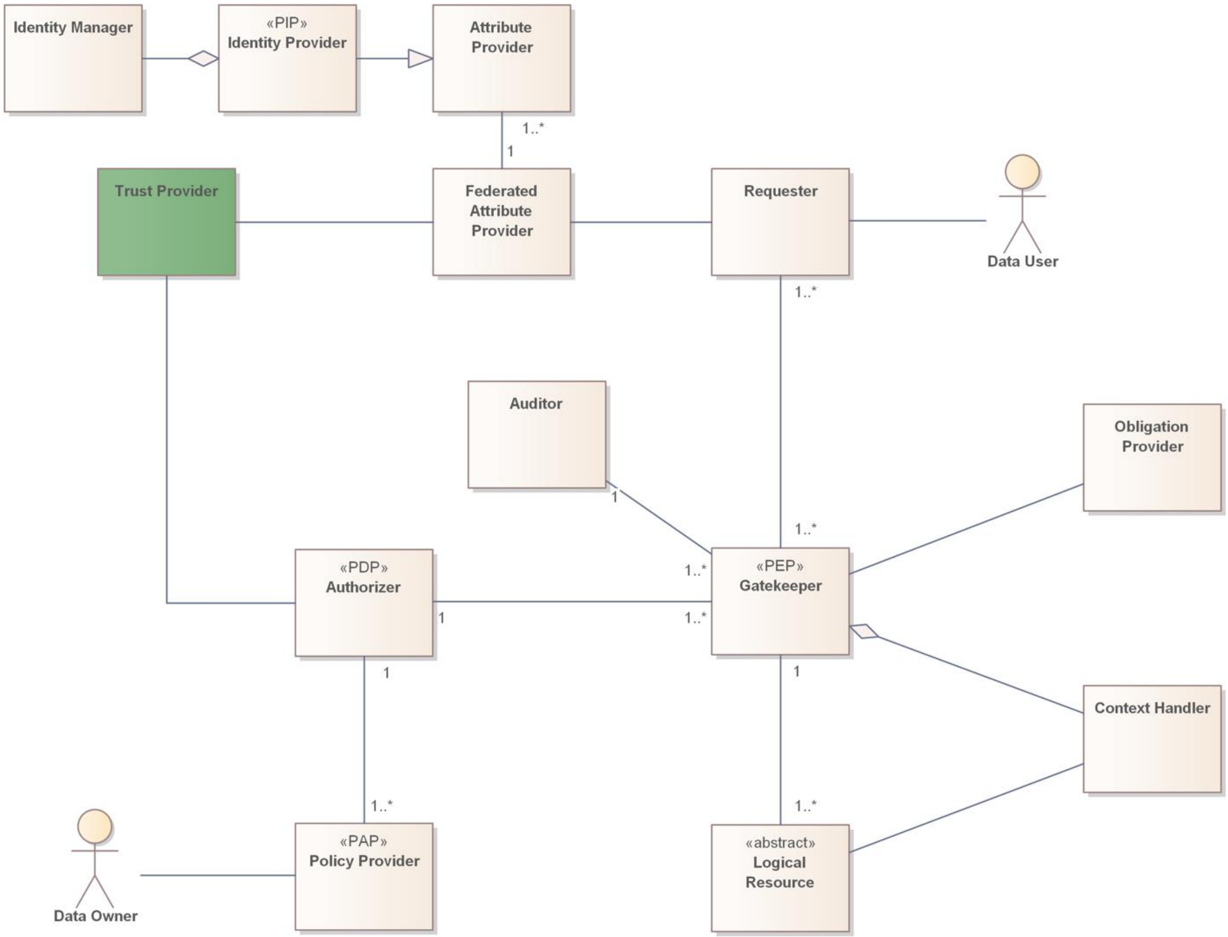


Figure 11 - Logical Architecture of GDDS DE Security Architecture

In the depicted architecture, the Trust Provider represents a macro-component which provides trust-related services and is contacted by components that need to obtain and/or verify trusted information (e.g., a Trust Provider might use a PKI for associating digital certificates to shared information or a Verifiable Credentials-based technology).

Finally, an Auditor is present to log all requested actions through the Gatekeeper, in order to enable monitoring-related functions (e.g., transaction metering, billing, etc.).

The logical architecture can be represented with the set of logical components in Figure 12. The Gatekeeper component acts as a proxy for GDDS DE interfaces, extending those interfaces with security information required by the access control framework.

It must be noted that in the presented logical architecture, the Policy Provider component is assumed to be made available by Data Owners. This means that each Data Owner should be able to provide a machine-readable description of its own data policy, possibly making the process of joining the GDDS DE more difficult. Besides, the Authorizer must be able to handle a variety of possible data policy formats and use them in its decision algorithm.

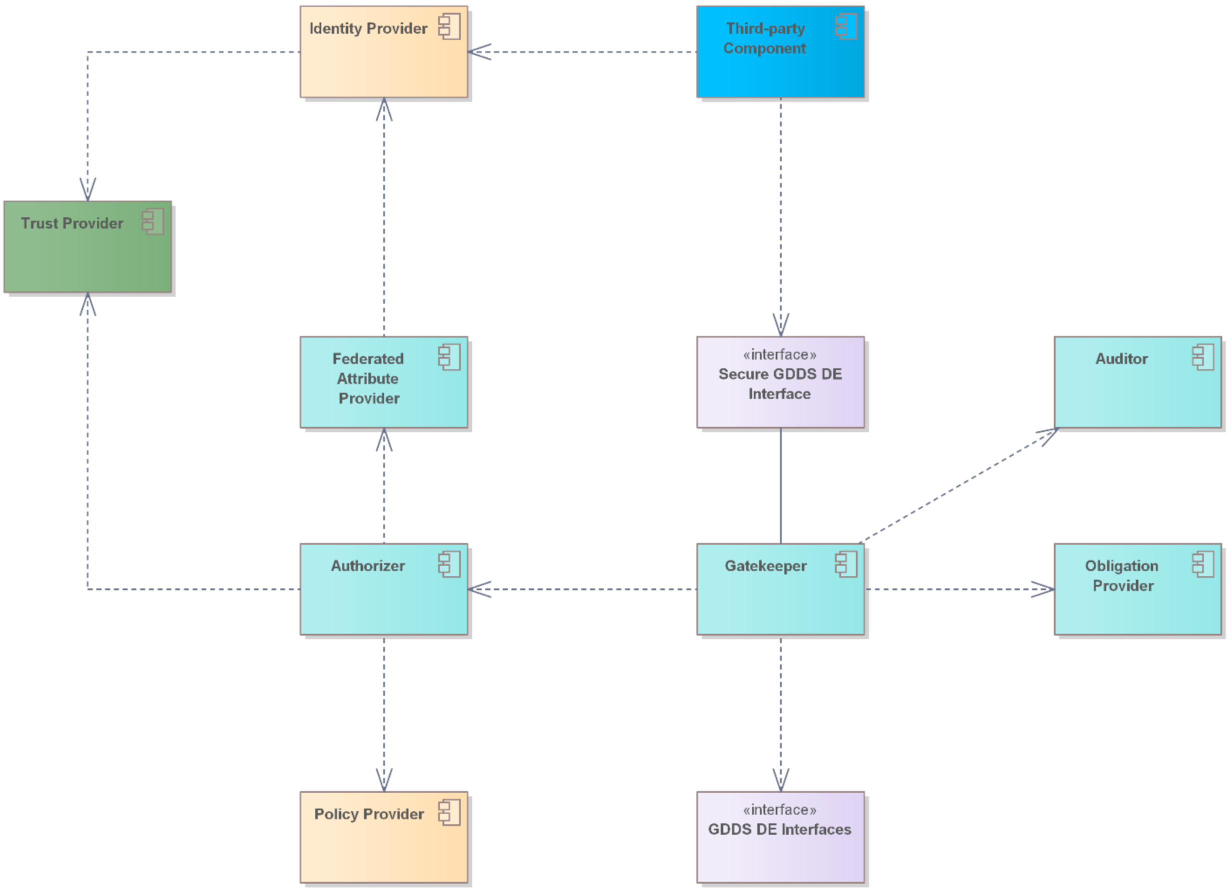


Figure 12 – Logical Components of the Security and Trust Architecture of the GDDS DE

5 GDDS DE Interoperability with Other Initiatives

In this section we are going to analyse the presented GDDS DE technical blueprint with respect to the Data Space Support Centre (DSSC) vision and other relevant initiatives. Consistency with the DSSC vision is necessary for the implementation of interoperable Common European Sectorial Data Spaces.

We will also analyse interoperability with the Destination Earth initiative, one of the major initiatives in the development of the European Green Deal; in particular, we will focus on the Destination Earth Data Lake.

Besides, we will consider the SIMPL and the DSBA documentation and analyse interoperability with data spaces developed on these frameworks.

Finally, we provide initial analysis of how the presented technical blueprint of the GDDS DE addresses the functional requirements identified in the “European Data Spaces - Scientific Insights into Data Sharing and Utilisation at Scale” report [1].

5.1 Data Space Support Centre Technical Blueprint

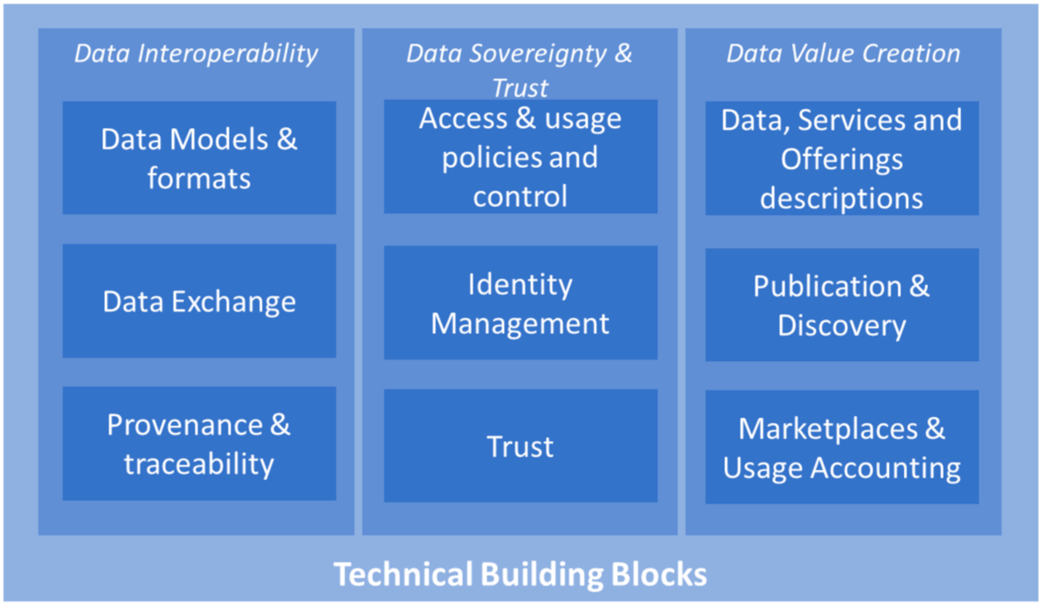
The Data Space Support Centre (DSSC) will release its first version of the technical blueprint in autumn 2023. However, the DSSC has already presented its high-level approach which clarifies that the DSSC blueprint will leverage the concept of building blocks defined in the Position Paper on “Design Principles for Data Spaces” [24]. These were further explored in the “Taxonomy of building blocks” document (draft version 0.5) which was shared with sectorial Data Spaces in June 2023. Therefore, at the moment, our analysis is carried out based on the content of this document.

For the scope of this document, we will focus on the technical building blocks from the DSSC taxonomy document (Figure 13), which are categorized as follows:

a) Data interoperability: building blocks to enable (semantic) interoperability between partners in a data space, enabling participants to specify their (domain specific and cross-domain) semantics, link them to (common) technical interfaces and record which data was exchanged with whom.

b) Data sovereignty and trust: building blocks to enable organisations to remain sovereign over their data, yet at the same time maintain trust in the data space as a whole. For example by ensuring access and usage control.

c) Data value creation: building blocks for creating value in a data spaces, e.g. by registering and discovering data offerings or services, providing marketplace functionality and enabling monetization of data and data services.



Data Interoperability	Data Sovereignty & Trust	Data Value Creation
Data Models & formats	Access & usage policies and control	Data, Services and Offerings descriptions
Data Exchange	Identity Management	Publication & Discovery
Provenance & traceability	Trust	Marketplaces & Usage Accounting

Technical Building Blocks

Figure 13 - Technical Building Blocks from DSSC Taxonomy Document

The building blocks in Figure 13 encapsulate high-level functionalities which were identified by the DSSC as necessary for the Data Spaces implementation and inter-Data

Space connection. It is important to note here that the design of the GDDS DE is based on the concept of Soft Infrastructure. This is further discussed in section 3.4, where the GDDS DE soft infrastructure is defined as comprised of the following two elements:

a) Agreements (including technical standards): these pertain to the governance sphere, which identifies the rules for participating in the GDDS DE.
b) Minimal set of (logical) components creating the digital environment: these components are in charge of providing the required interoperability solutions to connect the data consumers and data sources participating in the GDDS DE.

Therefore, in the following sub-sections we describe how the different building blocks are addressed by the logical components we described in section 4.

5.1.1 Data Models and Formats

Mainly two GDDS DE logical components address this building block: Dataset Transformer and Data Catalog. The Dataset Transformer is the component which is in charge of transforming the data which is shared in the GDDS DE. It provides all mediation, harmonization and transformation functionalities to enable the use of shared data in the cross-domain environment of the GDDS DE.

The Data Catalog is where the semantics of the shared data is captured, in the form of metadata elements describing the shared data from the different Data Sources. As recognized in section 4.4.3, for the Green Deal domain there is no common semantic service (thesaurus, ontology, etc.) which can be used across all the diverse domains.

5.1.2 Data Exchange

As in the case of the Data Models and Formats building block, the main GDDS DE logical component addressing this building block is the Dataset Transformer. In fact, it exposes a set of Uniform Data Access interfaces; each of these interfaces will comply with one standard recognized by the GDDS DE. This ensures interoperability with a wide range of existing Data Sources and Data Consumers, enabling also inter-data space interoperability.

Another GDDS DE component which contributes to this building block is the Data Mover, which allows Data Consumers to request the necessary data and make it available on a specific Cloud/HPC platform.

5.1.3 Provenance & Traceability

The minimum level of provenance is the information about the Data Source from which the data is shared. Besides, Data Sources can provide additional provenance information about the shared data in their metadata according to the specific standard they use. All provenance information will be captured in the metadata provided by the Data Catalog component.

Traceability is addressed by the Auditor component which logs all requests in order to enable monitoring-related functions (e.g., transaction metering, billing, etc.).

Finally, it is worth to note that the use of Persistent and Unique Identifiers (PUIs), and the correspondent PUI Provider and PUI resolver components, is key to enabling both provenance and traceability functionalities.

5.1.4 Access & usage policy and control

Access control is realized by applying the well-known and widely adopted XACML framework. This implements the so-called Remote Access Control approach, which refers to the protection of the digital content only during the access phase. Systems which realize this approach protect the digital content until this is transferred to the consumer. This choice is driven by the recognition that such an approach has a minor impact on GDDS DE participants, allowing in the initial phase an easier on-boarding. As recognized in section 4.8.1, the Remote Access Control approach can be seen as part of a wider end-to-end Digital Rights Management (DRM) system which can be introduced at a later stage.

In the proposed design, and in keeping with the data sovereignty principle, Data Owners can provide a machine readable description of their own data policy.

5.1.5 Identity Management

Identity Management is in charge of Identity Providers, that manage the entire lifecycle of user account management (creation, modification, suspension, etc.).

The GDDS DE technical blueprint was designed to support the use of different types of attributes, in addition to identity, to pass the access control of the XACML Policy Enforcement Point are provided by specific Attribute Providers. The Federated Attribute Provider is tasked with mapping attributes from the different Attribute Providers to a common representation in GDDS DE. Among possible Attribute Providers, the Identity Provider provides the Data User's authentication proof and the related identity attributes.

5.1.6 Trust

Trust refers to ensuring that a claim (e.g., "the user with ID 'id1' is a non-commercial user") is true. Achieving trust in a context like the GDDS can be built on top of two pillars:

c) Technical: to be able to ensure (verify) that the claim is from a certain organization.

d) Governance: acknowledge an organization as trustworthy, including the possibility of having different levels of trustworthiness for different types of claims.

In this technical blueprint the trust is handled by the macro-component Trust Provider. This provides trust-related services and is contacted by components that need to obtain and/or verify trusted information. At the technical level, several solutions exist to provide the desired functionality (e.g., a Trust Provider might use a PKI for associating digital certificates to shared information or a Verifiable Credentials-based technology). It is important to note here that compatibility with DSSC and, in turn, other sectorial Data Spaces is key to build an inter-Data Space trusted environment underpinning the envisioned single digital market.

5.1.7 Data, Services and Offerings descriptions

The GDDS DE supports a variety of metadata models. These are captured in the Data Sources Registry (as far as data) and in the Computing Infrastructure Registry (as far as HPC/Cloud platforms).

The GDSS DE approach is to allow participants to provide/request metadata according to their desired metadata model and encoding format. The necessary mediation and harmonization functionalities are provided by the GDDS DE components. Besides, original metadata can be enriched at the GDDS DE level, e.g., with information about the status of availability of the correspondent Data Source (by Status Checker component) or via the Metadata Enhancer that allows Data Consumers to add additional information to specific metadata (e.g., feedback, fit-for-purpose, etc.).

5.1.8 Publication & Discovery

The metadata provided by the participants are utilized to populate the corresponding catalogs: Data Catalog and Computing Infrastructure Catalog.

They expose a set of discovery interfaces which can be used by Data Consumers to discover available resources in the GDDS DE.

5.1.9 Marketplaces & usage accountings

This initial version of the GDDS DE does not include a Marketplace. However, it might be included in the second version of the technical blueprint.

Usage accounting is implemented by the Auditor component which logs all requests in order to enable monitoring-related functions (e.g., transaction metering, billing, etc.).

5.2 Destination Earth Data Lake

This analysis is based on the “DestinE - System Framework - Data Lake - High Level Description & Architecture” document released in December 2022 [40].

The DestinE Data Lake (DEDL) is one of three macro-components of Destination Earth initiative, besides the DestinE Core Service Platform (DESP) and the DestinE Digital Twin Engine (DTE). DEDL “fulfils the storage and access requirements for any data that is offered to DestinE users. It provides users with seamless access to the datasets, regardless of data type and location. Furthermore, the DEDL supports near-data processing to maximize throughput and service scalability” [40]. Figure 14 depicts a high-level overview of DEDL system interfacing with its external entities.

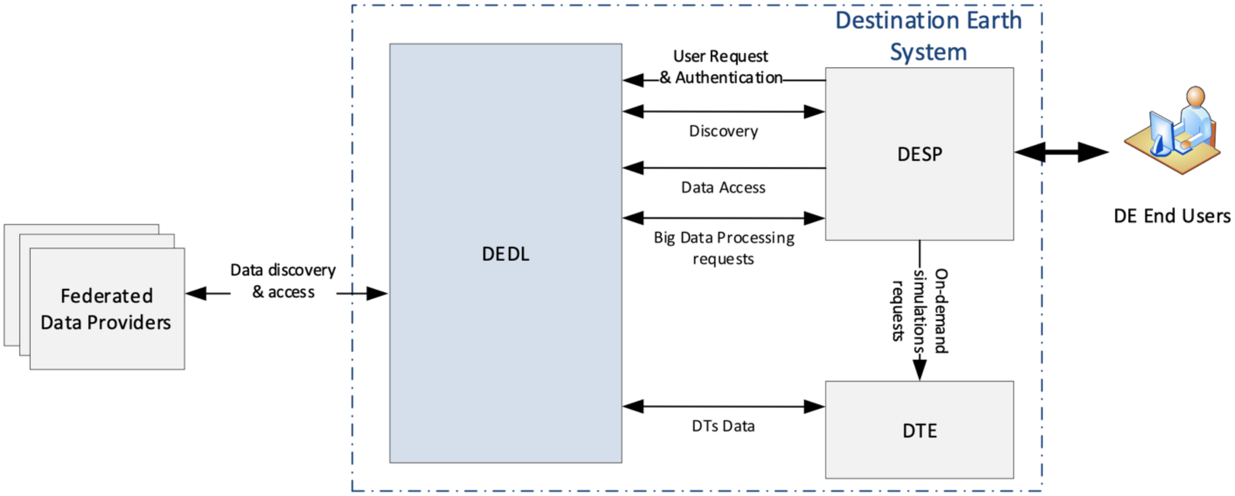


Figure 14 - DEDL System Context (source: DestinE - System Framework - Data Lake - High Level Description & Architecture)

Given its high-level features, the DEDL can interact with the GDDS DE both as a Data Source/Consumer and as a Computing Infrastructure provider.

The DEDL provides a set of Web APIs for discovery and access to its data. In particular, the architecture document mentions STAC and Opensearch as discovery interfaces. Data access is provided by DEDL via a Harmonized Data Access (HDA) layer, which "offers users a consistent, seamless access layer to a multitude of data pools, abstracting away the heterogeneous data access protocols. Besides, the DEDL HDA will allow users not only to directly access the discovered data but also to place data harvest requests and be notified when the request is completed, and the data is available for them" [40].

It is worth noting the conceptual alignment of the DEDL approach to data discovery and access with the one described in this technical blueprint of the GDDS DE. In fact, the GDDS DE logical components Data Catalog, Dataset Transformer and Data Mover provide essentially the same functionalities described above. Therefore, interoperability between GDDS DE and DEDL will be facilitated since each one of the two systems will be able to rely on each other's functionality for the execution of a discovery and access workflow.

In addition, DEDL will provide Web APIs for the exploitation of its computing infrastructure via its Big Data Processing unit which provides near-data processing capabilities inside the DEDL. To support different user needs the DEDL Big Data Processing portfolio offers three types of services [40]:

a) Applications/Environments - that are hosted on DEDL and with shared DEDL user access (e.g., JupyterHub, OpenDataCube's, DASK Gateway)

b) Functions (FaaS) - which can be executed by users from their applications

c) Infrastructure As A Service (Islet) - that users can use to deploy/run legacy applications or for purely data storage

While the Applications/Environments offerings are tailored for the internal use in the DEDL system, the IaaS (Islet) offering can be exploited by the GDDS DE. In fact, this kind of offering is already considered in the GDDS DE design. It is exploited by the Data Mover and the Data Processing Enabler components. Finally, the FaaS offering will require further investigation to better understand what kind of functions will be provided and how they can be modeled in the GDDS DE design.

5.3 SIMPL- Smart Middleware Platform (SMP)

In the preparatory phase of the Simpl procurement process, an architectural vision document [41] was released to describe the conceptual design of the Smart Middleware Platform (SMP).

The conceptual design of the SMP is built around the concept of the SMP Agent which provides common services on which data spaces can be built and enables interoperability between data spaces [41]. The SMP Agent is described as an abstract component that participants in a data space need to deploy to become part of the ecosystem.

It is important to note that the document recognizes that "the deployment of the SMP Agent in a data space can have various degrees of granularity" and "it is up to the single data space to decide how the SMP best provides value, and what level of granularity of the deployment fits best" [41].

Among the different deployment scenarios analyzed in the document, the one that fits best for the GDDS DE is the deployment of a single SMP Agent as gateway to interconnect the GDDS DE with SMP Agent-based data spaces. In fact, this will allow the use of the SMP Agent in a transparent way for all participants in the GDDS DE and at the same time enables the GDDS DE interoperability with other SMP Agent-based data spaces.

5.4 Data Space Business Alliance – Technical Convergence

In April 2023, the DSBA released the second version of its Technical Convergence Document [42] with the goal of achieving interoperability and portability of solutions across data spaces, by harmonizing technology components and other elements.

The high-level vision of the document is based on the concept of building blocks defined in Position Paper on "Design Principles for Data Spaces" [24]. In section 5.1 we have already analysed how the GDDS DE technical blueprint addresses the high-level functionalities encapsulated in the different building blocks.

The technical convergence document then focuses on a set of implementation solutions for the different functionalities. Although implementation details of the GDDS DE are out of the scope of this document, it is worth noting that the proposed GDDS DE design provides the necessary flexibility to accommodate interconnection with other data spaces based on different technological implementations. As an example, for the Data Models and Formats building block, DSBA is based on the use of data models defined by the Smart Data Models initiative which provides a library of data models for which the description and rendering in multiple data formats is provided. The Dataset Transformer of the GDDS

DE technical blueprint is the component that must provide the necessary transformation functionalities to/from such data models and format encodings.

5.5 JRC Report on European Data Spaces

The recently released "European Data Spaces - Scientific Insights into Data Sharing and Utilisation at Scale" report [1] identified a set of high-level requirements for the Common European Data Spaces. We provide in Table 5 an initial analysis of how the presented technical blueprint of the GDDS DE addresses the functional requirements identified in the report.

Table 5 -Analysis of High-level Requirements from JRC Science for Policy Report on European Data Spaces

Requirements	GDDS DE Logical Components
Data transfer and exchange. The core functionality of data spaces, enabling participants to transfer data to other participants.	Dataset Transformer Data Mover
Data publication and discovery. An effective mechanism for publication and discovery is expected to be a key functional requirement of data spaces, especially given the large amount of heterogeneous data expected to be made available in them.	Data Catalog Status Checker Metadata Enhancer
Data Storage. To support access to data, storage services can be either physical, i.e., based on independent copies of participants' data within the ecosystem, or virtual, providing access to data assets which are physically located in their owners' infrastructure.	Data Mover
Data interoperability. Features supporting the integration of heterogeneous data sources from the legal, organisational, technical and semantic perspectives.	Data Catalog Dataset Transformer

Data processing and analytics. The functionality of data spaces extends beyond making data available, and includes the utilisation of data for value-added applications, notably through data analytics and AI. Tools to streamline the development of AI solutions would be beneficial, especially if they target not only AI specialists but also domain-experts from the different sectors, e.g., through low-code, no-code, AutoML (automated machine learning methods and processes) and other approaches to make AI available for non-experts.	Data Mover Data Processing Enabler
Multi-tier support, federation and orchestration. Data spaces should provide develop- ment tools for multi-platform services that are supported by a wide range of underlying computing architectures, as well as interfaces for their orchestration – this is a key aspect of digital sovereignty.	Data Processing Enabler
Data pooling and collaboration. Collaboration tools are required to enable the joint development and exploitation of products and services by multiple participants in data spaces, possibly from different organisations and even economic sectors. Productivity and collaboration services could support and simplify the design, implementation and management of distributed processing workflows across ecosystem participants, ensuring an effective shared governance.	Not addressed at the moment
Identity, authentication and access control. These are key features upon which trust is built in the data sharing ecosystem,	Federated Attribute Provider XACML framework, based on data owners' data policy provisioning

enabling participants to control who can access their data assets.	
Privacy-preserving mechanisms. Ensuring data privacy is a key requirement for certain data spaces handling sensitive data (e.g., personally identifiable information or intellectual property). Data spaces should comply with the EU General Data Protection Regulation and provide data privacy features, such as anonymisation and masking services – they may in the future incorporate more advanced privacy-enhancing technologies, such as federated learning, secure multi-party computation and homomorphic encryption.	Features such as anonymization and masking can be implemented by the Obligation Provider.
Usage control policies. Building on access control functionality, additional features should enable participants in data spaces to determine not only who is allowed to access their data, but also the manner in which these data can be used, providing effective monitoring and enforcement functionality.	XACML framework implements a Remote Access Control approach, which can be seen as part of a wider end-to-end DRM which can be introduced at a later stage. See 4.8.1 for more details.
Compliance and auditing. This functional category encompasses features that enable participants in data spaces to attest and verify claims made by their peers regarding compliance with standards, regulations and general terms and conditions for using data and services. Such features include preconditions for making data available that are defined by their owners or by any other governing authorities.	Trust Provider (macro-component)
Transaction metering and billing. Features that enable participants in data spaces to monitor and monetise data	Auditor PUI Provider/Resolver

flows, as well as the consumption of their services within the ecosystem.	
Data governance. Data governance can be defined as the set of rules, policies, relations, decision-making structures and processes established among different kinds of actors to collect, share and use data. In general terms, it is understood as the correct management and maintenance of data assets and related aspects, such as data rights, data privacy, and data security, among others. While being a functional requirement on its own, data governance is also an essential prerequisite for many other (e.g., technical) functional requirements of data spaces. And in turn, the technologies used in a European data space should meet the requirements of data and information governance.	Data management functionalities are implemented by the different Data Providers participating in the GDDS DE. The Governance aspect is addressed by D4.1.
Data protection. Data spaces should protect the personal data of individuals that is shared within them, and comply with EU General Data Protection Regulation (GDPR) rules (European Union, 2016). The GDPR is a European law that establishes protections for privacy and security of personal data about individuals in European Economic Area (EEA)-based operations and certain non-EEA organizations that process personal data of individuals in the EEA. Privacy and data protection are also enshrined in the EU Treaties and in the EU Charter of Fundamental Rights.	All components will have to be implemented according to GDPR and other relevant legislations.

6 Conclusions and Inputs to Roadmap

This document described the first version of the Green Deal Data Space (GDDS) technical blueprint. Recognizing the need to design a solution which can evolve in the future responding to changes in the science/policy and technology contexts, the GDDS is based on the Digital Ecosystem (DE) paradigm. Such a paradigm fits particularly well with the vision of the Common European data spaces, and, specifically, the GDDS. In fact, this allows the GDDS to build on existing (and future) data systems, managed by organizations according to their own mandate and governance. Besides, it allows the GDDS to evolve in support of new applications that we cannot now imagine.

Therefore, the GDDS DE is designed as a Soft Infrastructure comprised of the following two elements:

a) Agreements (including technical standards): these pertain to the governance sphere, which identifies the rules for participating in the GDDS DE.
b) Minimal set of (logical) components creating the digital environment: these components are in charge of providing the required interoperability solutions to connect the data consumers and data sources participating in the GDDS DE.

The logical components are classified in two main categories: Core and Facilitators. The former identifies the logical components which are critical for the existence of the DE; the latter category identifies the components which facilitate the use of data available in the DE. Both Core and Facilitators components expose Web APIs which data consumer tools can use to exploit the GDDS DE data.

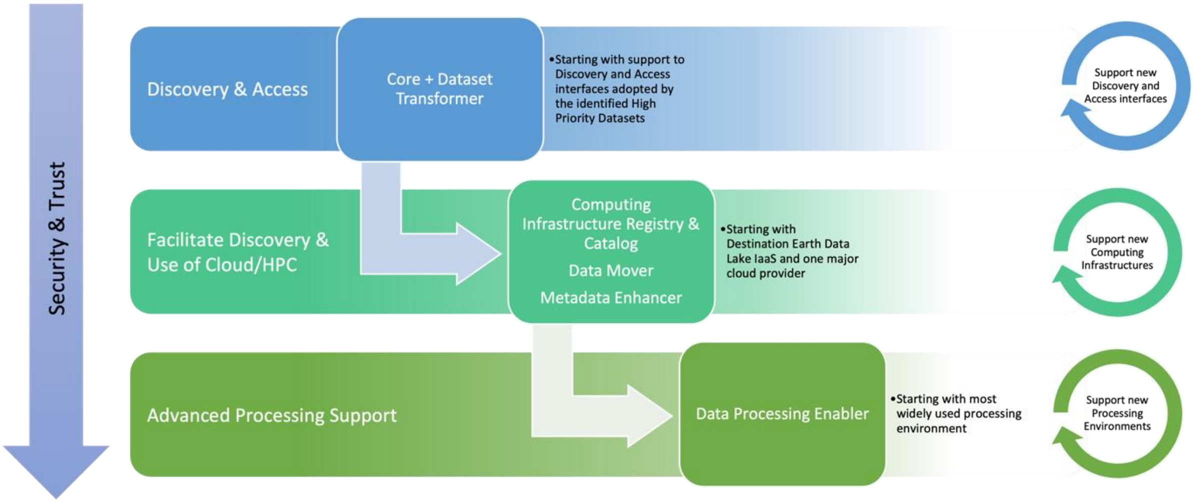

Security & Trust
Discovery & Access
Core + Dataset Transformer
•Starting with support to Discovery and Access interfaces adopted by the identified High Priority Datasets
Support new Discovery and Access interfaces
Facilitate Discovery & Use of Cloud/HPC
Computing Infrastructure Registry & Catalog
Data Mover
Metadata Enhancer
•Starting with Destination Earth Data Lake IaaS and one major cloud provider
Support new Computing Infrastructures
Advanced Processing Support
Data Processing Enabler
•Starting with most widely used processing environment
Support new Processing Environments

Figure 15 - Possible Development Roadmap of the GDDS DE

Based on the initial list of logical components described in this document, we identified a possible roadmap for their development, depicted in Figure 15.

We identified three high-level phases which support increasing levels of functionalities. For each phase we identified the logical components that provide the corresponding levels of functionalities and a possible initial set of supported systems.

The initial phase will address the basic functionalities of discovery and access, implementing the Core logical components and the Dataset Transformer. Initial implementation of these components will focus on supporting the Discovery and Access interfaces utilized by Data Sources providing the identified High Priority Datasets.

The second phase will target the facilitators which enable more advanced use of available data. In this phase it will be possible to exploit Cloud/HPC platforms capabilities to cope with Big Data requirements. The Computing Infrastructure Registry, Computing Infrastructure Catalog and the Data Mover components will enable data consumers to seamlessly move discovered data to the platforms where they operate. Initial implementation should support the Destination Earth Data Lake computing infrastructure and one major cloud provider. The Metadata Enhancer will be implemented in this phase as well, allowing to enrich descriptions of available data and therefore making the discovery phase more effective.

Finally, the third phase addresses advanced support for data processing, providing the implementation of the Data Processing Enabler. This will further facilitate the use of Cloud/HPC platforms, allowing data consumers to easily submit their algorithms implementation to different Cloud/HPC platforms. To this aim, the Data Processing Enabler will initially support the most widely used processing environments for scientific computation.

7 References

[1] E. Farrell *et al.*, “European Data Spaces - Scientific Insights into Data Sharing and Utilisation at Scale,” *JRC Publications Repository*, Jun. 12, 2023. https://publications.jrc.ec.europa.eu/repository/handle/JRC129900 (accessed Jul. 27, 2023).

[2] European Commission, Directorate-General for Communications Networks, and Content and Technology, *Shaping Europe’s digital future*. LU: Publications Office of the European Union, 2020. Accessed: Jul. 28, 2023. [Online]. Available: https://data.europa.eu/doi/10.2759/091014

[3] European Commission, “COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS 2030 Digital Compass: the European way for the Digital Decade.” 2021. Accessed: Jul. 26, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2021%3A118%3AFIN

[4] European Commission, “Commission Staff Working Document Annual Single Market Report 2023.” https://single-market-economy.ec.europa.eu/system/files/2023-01/ASMR%202023.pdf (accessed Jul. 26, 2023).

[5] European Commission, “A European strategy for data.” 2020. Accessed: Aug. 05, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC0066&from=EN

[6] European Commission, “Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act).” 2022. Accessed: Apr. 18, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R0868&from=EN

[7] European Commission, “Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on harmonised rules on fair access to and use of data (Data Act).” 2022. Accessed: Apr. 18, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0068&from=EN

[8] European Parliament, “Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast).” 2019. Accessed: Apr. 18, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L1024&from=EN

[9] European Commission, “Commission Implementing Regulation (EU) 2023/138 of 21 December 2022 laying down a list of specific high-value datasets and the arrangements for their publication and re-use.” 2023. Accessed: Apr. 18, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32023R0138&from=EN

[10] European Commission, “Commission Staff Working Document on Common European Data Spaces,” European Commission, SWD(2022) 45 final, Feb. 2022.

[11] European Commission, “DIGITAL EUROPE Work Programme 2021-2022.” 2021. Accessed: Apr. 19, 2023. [Online]. Available:

https://ec.europa.eu/newsroom/repository/document/2021-46/C_2021_7914_1_EN_annexe_acte_autonome_cp_part1_v3_x3qnsqH6g4B4JabSGBy9UatCRc8_81099.pdf

[12] GEO, "GEO Strategic Plan 2016-2025: Implementing GEOSS.," 2015. https://earthobservations.org/documents/ministerial/mexico_city/MS4_GEO%20Strategic%20Plan%202016-2025%20Implementing%20GEOSS_approved_by_GEO-XII.pdf (accessed Feb. 06, 2023).

[13] GEO, "GEO Engagement Strategy.," 2016. https://www.earthobservations.org/documents/geo_xiii/GEO-XIII-4-1_GEO%20Engagement%20Strategy.pdf (accessed Feb. 06, 2023).

[14] GEO, "Urban Resilience Engagement Priority.," 2020. https://earthobservations.org/documents/pb/me_202009/PB-18-07_Urban%20Resilience%20Engagement%20Priority.pdf (accessed Feb. 06, 2023).

[15] ISO, "ISO 19101-1:2014 Geographic information — Reference model — Part 1: Fundamentals." 2014.

[16] European Commission, "Consolidated text: Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)." 2019. Accessed: Apr. 19, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02007L0002-20190626&from=EN

[17] S. Nativi, M. Craglia, and J. Pearlman, "Earth Science Infrastructures Interoperability: The Brokering Approach," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 6, no. 3, pp. 1118–1129, Jun. 2013, doi: 10.1109/JSTARS.2013.2243113.

[18] S. Nativi, P. Mazzetti, M. Santoro, F. Papeschi, M. Craglia, and O. Ochiai, "Big Data challenges in building the Global Earth Observation System of Systems," *Environmental Modelling & Software*, vol. 68, pp. 1–26, Jun. 2015, doi: 10.1016/j.envsoft.2015.01.017.

[19] H. Guo *et al.*, "Big Earth Data science: an information framework for a sustainable planet," *null*, vol. 13, no. 7, pp. 743–767, Jul. 2020, doi: 10.1080/17538947.2020.1743785.

[20] S. Nativi and P. Mazzetti, "Geosciences Digital Ecosystems," in *Encyclopedia of Mathematical Geosciences*, B. S. Daya Sagar, Q. Cheng, J. McKinley, and F. Agterberg, Eds., Cham: Springer International Publishing, 2020, pp. 1–6. doi: 10.1007/978-3-030-26050-7_458-1.

[21] R. D. Blew, "On the Definition of Ecosystem," *Bulletin of the Ecological Society of America*, vol. 77, no. 3, pp. 171–173, 1996.

[22] S. Nativi, P. Mazzetti, and M. Craglia, "Digital Ecosystems for Developing Digital Twins of the Earth: The Destination Earth Case," *Remote Sensing*, vol. 13, no. 11, p. 2119, May 2021, doi: 10.3390/rs13112119.

[23] European Commission, "Preparatory actions for the Green Deal Data Space," 2021. https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/digital-2021-cloud-ai-01-prep-ds-green-deal (accessed Nov. 14, 2022).

[24] Nagel, Lars and Lycklama, Douwe, “Design Principles for Data Spaces - Position Paper,” Zenodo, Jul. 2021. doi: 10.5281/ZENODO.5105744.

[25] IEEE/ISO/IEC, “IEEE/ISO/IEC International Standard for Software, systems and enterprise--Architecture description,” IEEE, 2022. doi: 10.1109/IEEESTD.2022.9938446.

[26] N. Rozanski and E. Woods, *Software Systems Architecture: Working with Stakeholders Using Viewpoints and Perspectives*. Addison-Wesley, 2011.

[27] ISO/IEC, “ISO/IEC 10746-1:1998 Information technology — Open Distributed Processing — Reference model: Overview.” 1998.

[28] H.-W. Hilse and J. Kothe, *Implementing persistent identifiers: overview of concepts, guidelines and recommendations*. London: Consortium of European Research Libraries, 2006.

[29] P. Mell and T. Grance, “The NIST Definition of Cloud Computing,” p. 7, 2011.

[30] E. Boldrini, S. Nativi, S. Pecora, I. Chernov, and P. Mazzetti, “Multi-scale hydrological system-of-systems realized through WHOS: the brokering framework,” *International Journal of Digital Earth*, vol. 15, no. 1, pp. 1259–1289, Dec. 2022, doi: 10.1080/17538947.2022.2099591.

[31] U. Schindler and M. Diepenbroek, “Generic XML-based framework for metadata portals,” *Computers & Geosciences*, vol. 34, no. 12, pp. 1947–1955, Dec. 2008, doi: 10.1016/j.cageo.2008.02.023.

[32] EUDAT, “B2FIND,” *EUDAT*, Mar. 01, 2015. https://eudat.eu/services/b2find (accessed Oct. 08, 2021).

[33] R. Petrie *et al.*, “Coordinating an operational data distribution network for CMIP6 data,” *Geoscientific Model Development*, vol. 14, no. 1, pp. 629–644, Jan. 2021, doi: 10.5194/gmd-14-629-2021.

[34] M. Santoro, P. Mazzetti, and S. Nativi, “The VLab Framework: An Orchestrator Component to Support Data to Knowledge Transition,” *Remote Sensing*, vol. 12, no. 11, Art. no. 11, Jan. 2020, doi: 10.3390/rs12111795.

[35] M. Santoro, P. Mazzetti, and S. Nativi, “Virtual earth cloud: a multi-cloud framework for enabling geosciences digital ecosystems,” *International Journal of Digital Earth*, vol. 16, no. 1, pp. 43–65, Dec. 2023, doi: 10.1080/17538947.2022.2162986.

[36] ITU, “X.800 : Security architecture for Open Systems Interconnection for CCITT applications.” 1991. [Online]. Available: https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-X.800-199103-I!!PDF-E&type=items

[37] R. W. Shirey, “Internet Security Glossary, Version 2,” Internet Engineering Task Force, Request for Comments RFC 4949, Aug. 2007. doi: 10.17487/RFC4949.

[38] B. Y. Fraser, “Site Security Handbook,” Internet Engineering Task Force, Request for Comments RFC 2196, Sep. 1997. doi: 10.17487/RFC2196.

[39] OASIS, “eXtensible Access Control Markup Language (XACML) Version 3.0,” 2013. https://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html (accessed Jun. 29, 2023).

[40] EUMETSAT, “DestinE - System Framework - Data Lake - High Level Description

& Architecture.” 2022.
[41] Simpl, “Architecture Vision Document.” 2023. [Online]. Available: https://ec.europa.eu/newsroom/dae/redirection/document/86241
[42] DSBA, “Technical Convergence.” 2023. Accessed: Jul. 25, 2023. [Online]. Available: https://data-spaces-business-alliance.eu/wp-content/uploads/dlm_uploads/Data-Spaces-Business-Alliance-Technical-Convergence-V2.pdf

Annex A: Possible Service Interfaces and Data/Metadata Models to be supported in GDDS DE

Table 6 lists some possible relevant service interfaces and metadata/data models for discovery and access to be supported by the GDDS DE. The list is composed based on the experience of other large multidisciplinary data sharing initiatives (e.g., the Global Earth Observation System of Systems, GEOSS).

Table 6 - List of possible relevant service interfaces and metadata/data models for discovery and access

Name	Brief Description
OGC WCS 1.0, 1.1, 1.1.2	Discovery (coverages inventory) and access interfaces
OGC WMS 1.3.0, 1.1.1	Discovery (maps inventory) and access interfaces
OGC WFS 1.0.0	Discovery (features inventory) and access interfaces
OGC WPS 1.0.0	Discovery (processes inventory) and access interfaces
OGC SOS 1.0.0	Discovery (sensors inventory) and access interfaces
OGC CSW 2.0.2 Core, AP ISO 1.0, ebRIM/CIM, ebRIM/EO, CWIC	Discovery interface and metadata profiles
HDF	Metadata and data encoding
HMA CSW 2.0.2 ebRIM/CIM	Discovery interface
GeoNetwork (versions 2.2.0 and 2.4.1) catalog service	Discovery interface
Deegree (version 2.2) catalog service	Discovery interface

ESRI ArcGIS Geoportal (version 10) catalog service	Discovery interface
WAF Web Accessible Folders 1.0	Discovery and access interfaces and metadata model
FTP - File Transfer Protocol services populated with supported metadata	Discovery and access interfaces
THREDDS 1.0.1, 1.0.2	Discovery and access interfaces
THREDDS-NCISO 1.0.1, 1.0.2	Discovery and access interfaces, and metadata model
THREDDS-NCISO-PLUS 1.0.1, 1.0.2	Discovery and access interfaces, and metadata model
CDI 1.04, 1.3, 1.4 1.6	Discovery interface and metadata model
GBIF	Discovery and access interfaces, and metadata model
OpenSearch 1.1 accessor	Discovery interface
OAI-PMH 2.0 (support to ISO19139 and dublin core formats)	Discovery interface and metadata model
NetCDF-CF 1.4	Metadata and data model
NCML-CF	Metadata and data model
NCML-OD	Metadata and data model
ISO19115-2	Metadata model
GeoRSS 2.0	Access interface, and metadata model
GDACS	Access interface, metadata and data models

DIF	Metadata and data model
SITAD (Sistema Informativo Territoriale Ambientale Diffuso) accessor	Discovery and access interfaces
INPE	Discovery and access interfaces
HYDRO	Discovery and access interfaces
EGASKRO	Discovery and access interfaces
RASAQM	Discovery and access interfaces
IRIS event	Discovery and access interfaces, metadata model
IRIS station	Discovery and access interfaces, metadata model
UNAVCO	Discovery and access interfaces, metadata model
KISTERS Web - Environment of Canada	Discovery and access interfaces
DCAT	Discovery interface and metadata model
CKAN	Discovery interface and metadata model
HYRAX THREDDS SERVER 1.9	Discovery and access interfaces
Socrata Open Data API	Data discovery service
ESRI shapefile	File format
.KML	File format
GML	File format
GeoJSON	File format

TIFF GeoTIFF	File format

Annex B: Legal and Ethical Assessment Methodology

The Legal and Ethical Assessment Methodology provided by the Ethics Advisor of the GREAT project, serves as a comprehensive framework designed to systematically identify, evaluate, and address legal and ethical risks associated with a project's deliverables. Following a "by design" approach, this methodology is seamlessly integrated into the project's technical workflow, ensuring the consideration of legal and ethical aspects throughout the project's lifecycle. Its primary objectives encompass optimizing technical and business goals, ensuring compliance with relevant legal standards and ethical principles, and fostering ongoing competence-building within the research community involved.

Implemented in three key steps, the methodology begins with a preliminary meeting involving Work Package (WP) leaders, where the foundational literature and guiding legal and ethical principles are presented. The checklist analysis phase follows, employing a proactive "learning-by-doing" approach to identify potential gaps and risks across domains such as Data Privacy, Ownership, Licenses, Competition, Artificial Intelligence, and Social Media. Feedback from the Ethics Advisor on identified gaps and risks is integrated into the final deliverable, concurrently nurturing the skills necessary for crafting resilient legal and ethical solutions. These solutions address a breadth of domains and prioritize the overall impact of the deliverable while aligning with research and business goals, fostering a comprehensive legal and ethical framework.