



EuroGEO Showcases: Applications Powered by Europe

D3.8: E-SHAPE GUIDE DEVELOPMENT

“BEST PRACTICES”



The e-shape project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 820852



(blank page)

ABSTRACT

The e-shape H2020 Project brings together decades of public investment in Earth Observation and more recently in cloud capabilities into services to the citizens, the industry, the decision-makers, and researchers. e-shape promotes the development and uptake of 37 cloud-based pilot applications, addressing the Sustainable Development Goals, The Paris Agreement, and the Sendai Framework.

The pilots address seven different thematic domains, building on GEOSS and on the Copernicus data pool as well as diverse computational infrastructures. The consortium gathers 68 members from 17 European countries, Ethiopia, Egypt, and Israel. It is a major European contribution to EuroGEO.

This Guide for European Earth Observation application developers, decision-makers, and experts, delivering best practices to use Earth Observation resources is based on the experience collected during the project and shares such knowledge in an accessible way. It provides a unique source and guidelines to increase the usage and exploitation of Earth Observation in the thematic domains addressed by e-shape. It aligns as much as possible with the European Commission's digital strategy supporting "the digital transformation journey broadening the scope to the move from IT to digital transformation, from digital or EO scientific skills to EO digital culture and from technology as a service provider to digitalization and digital ready policymaking." This is done by mixing digital background, scientific challenges and policy-making support to make them accessible and shareable in a common baseline of understanding that is, maybe, the essence of an Earth Observation "digital culture".

e-shape has captured the requirements and lessons learned out of the implementation of the 37 pilots over more than 70 platforms. It identified all essential elements to develop a successful Earth Observation application that builds on top of the available European Earth Observation resources and how to publish its results to make the Findable, Accessible, Interoperable and Reusable on the web.

e-shape has generated a large amount of complex information and a major challenge of the Guide is to address the concerns in a progressive, logical, and comprehensive way, making complex technical issues and challenges accessible to all. Sharing this baseline of knowledge and understanding will enable further, broader and faster collaborations.

The e-shape guide for development does not attempt to develop expertise on all topics but rather to share a holistic approach and understanding, to support all the EO stakeholders' profiles in their global process from data to product delivery and in their interactions with their multidisciplinary teams. Its goal is to mainstream a baseline of knowledge to develop the community, intensify the connections between the different expertise and skills, and speed up the development process from the concept to a prototype and operations in a domain that constantly evolves, and is resource and knowledge intensive.

The concept is to summarize succinctly the different topics and exemplify them as often as possible with the experience of selected e-shape pilot. The value is less in the technical details than in the holistic approach from an idea to results exploitation, taking the best of each e-shape pilot, whatever their thematic domain to support cross domain and cross expertise fertilization.

The wealth of knowledge collected is organized via an abstract generic and reproducible workflow from data discovery to results publication, from the simplest to the most complex issues. The report is articulated along the steps listed in Figure 1.

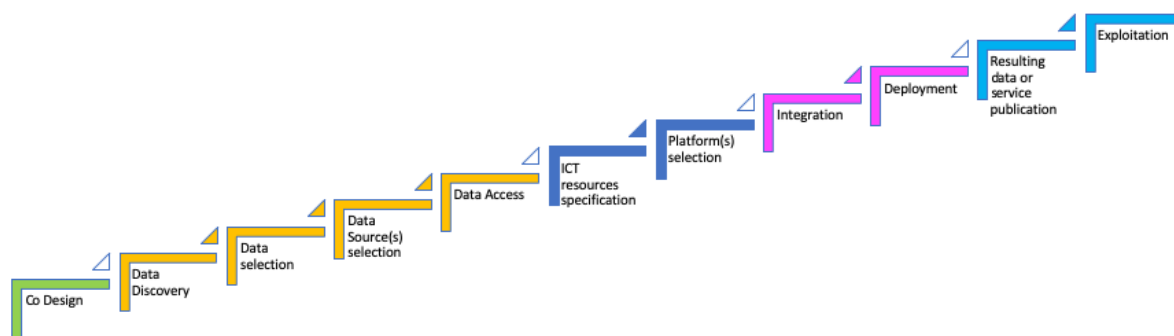


Figure 1: Generic reproducible Pilot development workflow. This scheme defines the structure of this report.

This approach is based on 11 steps, inspired by the proven and successful 5-step-user experience defined by the NextGEOSS H2020 project, that has been extended and detailed to structure and organize the topics identified by the pilots in the e-shape initial assessment as issues or challenges. The NextGEOSS 5-step-user experience was itself an adaptation of Terradue corporate Platform Operations proved procedures for the Ellip Solutions¹.

Access e-shape's Pilots: <https://tinyurl.com/e-shape-Pilots>

Other access:

- <https://e-shape.eu/index.php/all-pilots>
- <https://tinyurl.com/GEO-Knowledge-Hub>
- <https://tinyurl.com/GEO-DAB>

¹ References:

- NextGEOSS 5-step user experience:
https://ceos.org/document_management/Working_Groups/WGISS/Meetings/WGISS-46/3.%20Wednesday/2018.10.24_13.30_NextGEOSS.pdf
- Ellip solutions:
https://www.earthobservations.org/uploads/wp23_25_global_wildfire_information_system_implementation_plan.pdf



DOCUMENT TYPE	Report
DOCUMENT NAME:	D3.8 e-shape Guide development – “Best Practices “
VERSION:	Final
DATE:	May 31st, 2023
STATUS:	FINAL
DISSEMINATION LEVEL:	PU

AUTHORS, REVIEWERS			
AUTHOR(S):	Marie-Françoise Voidrot		
AFFILIATION(S):	OGCB		
FURTHER AUTHORS:			
PEER REVIEWERS:	PMT		
REVIEW APPROVAL:	PMT		
REMARKS / IMPROVEMENTS:			

VERSION HISTORY (PRELIMINARY)			
VERSION:	DATE:	COMMENTS, CHANGES, STATUS:	PERSON(S) / ORGANISATION SHORT NAME:
VO.1	10 April 2023	1 st draft version	Marie-Françoise Voidrot (OGCE); Alexia Tsouni (NOA); Andrea Vajda (FMI); Andy Nelson (UTWENTE); Annelies Hommersom (Water insight); Antonio Parodi (CIMA F.); Eleni Athanasopoulou (NOA); Francesca Piatto (EARSC); Kees de Bie (UTWENTE); Kostis Tsapraillis (NOA); Laurent Tits (VITO); Lionel Menard (ARMINES); Maja Zuvela Aloise (ZAMG); Marco Folegani (MEE0); Marion Schroedter-Homscheidt (DLR); Mikko Strahlendorff (FMI); Miriam Kosmale (FMI); Nuno Grosso (DEIMOS); Otto Hyvärinen (FMI); Panagiotis Kosmopoulos (NOA); Pablo Ezquerro Martin (EuroGEOSurveys); Patricia Gasparf (IPMA); Piotr Zaborozski (OGCE); Saskia Buchholz (DWD); Sergio Cinnirella (CNR); Ulf Mallast (UFZ)
VO.2	30 April 2023	<ul style="list-style-type: none"> Addition of example for AI use Update of the paragraph on Licenses 	Marie-Françoise Voidrot
V1.0		1 st draft version reviewed	PMT
V2.0		Implementation of review comments	Marie-Françoise Voidrot
vfinal		Final version for submission to the EC	PMT

VERSION NUMBERING	
v0.x	draft before peer-review approval
v1.x	After the first review
v2.x	After the second review
Vfinal	Deliverable ready to be submitted

STATUS / DISSEMINATION LEVEL			
STATUS		DISSEMINATION LEVEL	
S0	Approved/Released/Ready to be submitted	PU	Public
S1	Reviewed	CO	Confidential, restricted under conditions set out in the Grant Agreement
S2	Pending for review		
S3	Draft for comments	CI	Classified, information as referred to in Commission Decision 2001/844/EC.
S4	Under preparation		

TABLE OF CONTENT

<i>Abstract</i>	3
<i>Table of content</i>	8
<i>1 Introduction</i>	14
<i>2 Co-Design</i>	16
2.1 Introduction	16
2.2 On-going reflection on further co-design routinization	23
<i>3 Data, Process, and Applications discovery</i>	25
3.1 Introduction	25
3.2 From data download then process to data processing via online applications then download	28
3.3 Community Portal, GEOSS, Google Data Search	30
3.4 Standards for Data Discovery	32
<i>4 Data Selection</i>	34
4.1 Introduction	34
4.2 Data assets analysis	34
4.3 Data quality	49
<i>5 Data source(s) selection</i>	60
5.1 Introduction	60
5.2 Data Cubes Data organization and services as an enabler for efficient exploration of multidimensional data	60
5.3 Umbrella Sentinel Access Point (USAP) as an enabler for data access efficiency and resilience	63
5.4 General considerations	64
<i>6 Data access</i>	65
6.1 Introduction	65
6.2 Data Privacy and security	65
6.3 Technical limitations	67
6.4 Legal and regulatory requirements	68
6.5 Data complexity	68
6.6 User interaction: accessing the Data or running the application	69
6.7 Login service	69
6.8 Earth Observation Data portals	69
6.9 Application Programming Interfaces (APIs)	69
6.10 Jupyter Notebook	70
<i>7 New paradigm of Digital Twins</i>	70
7.1 New Processing Technologies: Artificial Intelligence, Machine Learning, Deep Learning	73
7.2 Cloud technologies for Earth Observation	74
7.3 Synthesis on the usability of Cloud Technologies for Earth Observations - Theory and Practice Status	77

8	<i>Information and Communication Technology- ICT- resources specification</i>	79
8.1	A change of paradigm	79
8.2	Sponsorship for Cloud resources	80
9	<i>Platform selection</i>	81
9.1	Introduction	81
9.2	Cloud or High-Performance Computing (HPC)?	82
9.3	Selecting a European Platform as a Service Cloud platform	83
10	<i>Integration</i>	88
10.1	Open source tools	88
10.2	Containers	88
11	<i>Deployment</i>	90
12	<i>Resulting in data or service publication or dissemination</i>	92
12.1	Introduction	92
12.2	Publication platforms overview	93
12.3	Data and service metadata	94
12.4	Defining a data license	94
12.5	Publishing on the web via data portals	98
12.6	Making data accessible via FTP (File Transfer Protocol)	110
12.7	Disseminating data via Satellite dissemination	110
12.8	Publication standards	111
13	<i>Exploitation</i>	115
13.1	Common terminology ambiguities	115
13.2	Describing the fitness for use and limits of operational products	116
13.3	Managing input data changes	116
13.4	Single point of failure analysis	119
13.5	Using Web Analytics tools or services to optimize the publication	119
13.6	Reproducibility	121
13.7	Data Management Plan	122
13.8	Quotation and Billing	125
13.9	The value of Standards, Data Models and Best Practices	125
13.10	Standards Compliance	126
13.11	Coherence and compliance of e-shape GEO DMP regarding the INSPIRE directive	127
13.12	Pilot Exploitation Readiness Level - PERL	130
14	<i>Annexes short description</i>	132
14.1	Annex 1: e-shape pilot applications	132
14.2	Annex 2: Glossary	132
14.3	Annex 3: Copernicus Services used by the e-shape pilots	132
14.4	Annex 4: Open source software and packages used by the e-shape pilots	132

14.5	Annex 5: Standards used by the e-shape pilots	132
14.6	Annex 6: Platforms used by the e-shape pilots	132

FIGURES

Figure 1: Generic reproducible Pilot development workflow. This scheme defines the structure of this report.	4
Figure 2: How do the user discover, access or run the pilot and how does the pilot interact with EO resources?	15
Figure 3: Co Design activities in the Development workflow.....	16
Figure 4: Representation of the "data journey" for the targeted state based on the data-information-usage framework	18
Figure 5: Anonymized data-information-usage framework completed using the answers to the initial assessment questionnaire.....	20
Figure 6: Graph synthesizing co-design type 1 outcomes in a 'resilient-fit' perspective.	22
Figure 7: Data Discovery in the Development workflow	25
Figure 8:User-Pilot interaction	26
Figure 9: Data Selection in the Development workflow.....	34
Figure 10 : Global Observing system (source WMO: https://public.wmo.int/en/programmes/global-observing-system).....	35
Figure 11: Copernicus Services, a user driven approach.....	47
Figure 12:INSPIRE Technical Guidelines use ISO 19157 Geographic Information Data quality. Source: EEA.	50
Figure 13 : Prediction skills for temperature (MSESS).....	52
Figure 14 : Comparison of “Downscaling of climate scenarios using urban climate model 100x100 indices” with SPARTACUS/OKTAV climate scenarios for Austria.....	53
Figure 15 : Snow Depth and T 2m Reforecast adjusted with statistical methods	55
Figure 16 : Land Use refinement with local municipal authority data	57
Figure 17 : Land Use refinement with Land Information System Austria (LISA) 1m data	58
Figure 18 : Past observed (1961–2019) and future projected (5-year running mean regional climate model simulations for scenarios RCP2.6, RCP4.5, and RCP8.5 in the period 1970–2100) annual mean temperatures for Austria (left) and climate change signal compared to the 1971–2000 period (right).	58
Figure 19: Data source Selection in the Development workflow	60
Figure 20: Different options for Data Cubes implementations	61
Figure 21: Data access in the Development workflow	65
Figure 22: e-shape contribution to a SWOT analysis of Cloud technologies for Earth Observations	74
Figure 23: ICT Resources specifications in the Development workflow	79
Figure 24:Platform Selection in the Development workflow	81
Figure 25: Progress of Earth Observation Technologies in Europe in the last 10 years	83
Figure 26: Major European platforms capacities evolution	84
Figure 27: Integration in the Development workflow	88
Figure 28: Deployment in the Development workflow	90
Figure 29: Resulting Data or service publication in the Development workflow	92



Figure 30:Inputs and outputs Licenses	95
Figure 31: NextGEOSS data portal	106
Figure 32: Ticket registration over NextGEOSS Service Desk.....	107
Figure 33: CREODIAS third party services catalog	108
Figure 34: Exploitation in the Development workflow.....	115
Figure 35 HarvesterSeasons.com usage as analysed by Google Analytics	120
Figure 36: OGC Standard Architecture Diagram.....	126
Figure 37 : Metadata record validation: Step #1 screenshot.....	128
Figure 38 : Metadata record validation: Step #2	129
Figure 39 : Metadata record validation: Step #3	129

TABLES

Table 1 : Classification of co-design needs to grow an ecosystem of efficient service	17
Table 2: Distinction between 'quick-fit' and 'resilient-fit' perspectives for the 4 types of co-design	19
Table 3: Main characteristics of the available SAR satellites and constellations.....	36
Table 4: Publication platforms details.....	99



BOXES AND e-shape's PILOTS TESTIMONIALS

Box 2-1: Co-design. Lessons learnt from Showcase 5: Water resources management Pilot 6 EO based phytoplankton biomass for WFD reporting (WI).....	24
Box 3-1: Data discovery. Lessons learnt from the mySITE Pilot MyEcosystem showcase.....	28
Box 3-2 : Data download to processing. Lessons learnt from myVARIABLE Pilot MyEcosystem showcase	29
Box 3-3: Data download to processing. Lessons learnt from Showcase 4 Water resources management ,Pilot 5: Sargassum detection for seasonal planning	30
Box 3-4: Community portals et al. Lessons learnt from Pilot Forestry conditions (more efficient forestry operations with lower environmental impact and carbon emissions) from the Climate showcase	31
Box 3-5:Community portals et al. Lessons learnt from the mySITE (data provision, visualisation tools, and ecosystem status indicators) Pilot MyEcosystem showcase	32
Box 4-1: Satellites used by Pilots. Lessons learnt from the Showcase MyEcosystem pilot mySPACE (better monitoring climate drivers in 25 protected areas).....	38
Box 4-2: In situ data. Lessons learnt from the Showcase MyEcosystem	39
Box 4-3: Citizen Data Science. Lessons learnt from the GEOGLAM Pilot - The Food Security and Sustainable Agriculture showcase example	40
Box 4-4: Numerical models Lessons learnt from the EO-based surveillance of Mercury pollution (Minamata Convention) Pilot - Health Surveillance showcase	41
Box 4-5: Numerical models. Lessons learnt from FYWA - Early WARNING System for Mosquito-Borne Diseases Pilot - Health Surveillance showcase.....	41
Box 4-6: Numerical models. Lessons learnt from the Health Surveillance Air Quality Pilot - Health Surveillance showcase.....	42
Box 4-7: Sub-seasonal forecasts for seasonal preparedness in extreme climates. Lessons learnt by FMI.....	44
Box 4-8: The need for consistent time series - an example from the Vegetation-Index Crop-Insurance in Ethiopia pilot in the Food Security and Sustainable Agriculture showcase_	45
Box 4-9: “ARD online” as an enabler of javascript-based EO Apps	46
Box 4-10: Essential Variables. Testimonial from Showcase MyEcosystem Pilot myVARIABLE (MLU, UT, SYKE, WR)	49
Box 4-11: Data quality. Lessons learnt by FMI, ZAMG/Geosphere, DWD across multiple Pilots	50
Box 4-12: Data quality. Lessons learnt by DWD and ZAMG/Geosphere, in the Climate Showcase.....	52
Box 4-13: Data quality. Lessons learnt from FMI, DWD and ZAMG across several Pilots.	54
Box 5-1: Data Cubes. Some Data cubes used by e-shape pilots.....	61
Box 5-2: A single point of access for sentinel data, connecting multiple data sources.....	64
Box 6-1:Data privacy & security. Lessons learnt from Showcase 1: Food Security and Sustainable Agriculture Pilot3: Vegetation-Index Crop-Insurance in Ethiopia (drought insurance for smallholder farmers) Pilot	66
Box 6-2: Data privacy & security. Lessons learnt from the Water resources management showcase, the development of Pilot 5 Monitoring fishing activity	66
Box 7-1: Digital twins. Lessons learnt from MyEcosystem Pilot 2: mySITE (data provision, visualisation tools and ecosystem status indicators).....	72
Box 7-2: New processing technologies. Lessons learnt from Showcase 1: Food Security and Sustainable Agriculture Pilot 1: GEOGLAM	73
Box 8-1: ICT resources specifications. Lessons learnt from Showcase 4: MyEcosystem Pilot 1: mySPACE (better monitoring climate drivers in 25 protected areas.	80



Box 9-1: Cloud platform selection. Lessons learnt from Solar Energy nowcasting and short-term forecasting system (management support for solar energy plant operators) Pilot Renewable Energy Showcase Example	86
Box 9-2: Feedback on Using a DIAS Platform for processing of EO data. Lessons learnt from Showcase 5: Water resources management Pilot 3: Diver Information on Visibility in Europe (coastal water quality monitoring).	87
Box 10-1: Using Dockers to develop an API for DIVE. Lessons learnt from Showcase 5: Water resources management Pilot 3: Diver Information on Visibility in Europe (coastal water quality monitoring)	89
Box 12-1: EO-Based Surveillance of Mercury Pollution Services and data resources discovery and access via the GEOSS DAB catalogue after direct publication of the Pilot's outcomes	103
Box 12-2 High photovoltaic penetration at urban scale pilot's metadata details in the GEOSS DAB after harvesting of the metadata from the GeoNetwork implementation from the WebService Energy	104
Box 12-3: Publishing on the GEO Knowledge Hub. Lessons learnt from EO-based surveillance of Mercury pollution (Minamata Convention) Pilot - Health Surveillance showcase example.	105
Box 12-4: GeoNetcast. Lessons learnt from the agriculture VICI - Vegetation-Index Crop-Insurance in Ethiopia pilot	111
Box 12-5: EUMETCast. Lessons learnt from Pilot2: High PV penetration in urban area (economic opportunities for solar energy through urban solar mapping) of the Renewable Energy showcase	111
Box 13-1: Exploitation. Lessons learnt from the Pilot Diver Information on Visibility in Europe (coastal water quality monitoring) of the Water resources management showcase.	116
Box 13-2: Fitness for use. Lessons learnt from Assessing Geo-hazard Vulnerability of Cities and Critical Infrastructures Pilot Disasters Resilience Showcase Example.....	116
Box 13-3: Managing input data changes e-shape agriculture VICI - Vegetation-Index Crop-Insurance in Ethiopia pilot experience.....	117
Box 13-4: Single point of failure analysis. Lessons learnt from The Pilot 2: GEOSS for Disasters in Urban Environment (improved resilience of cities, infrastructure and ecosystems to disasters) of the Showcase Disasters Resilience	119
Box 13-5: Web analytics. Lessons learnt from the Forestry condition Pilot.	121
Box 13-6: Data Management Plan. Lessons learnt from Showcase 4: MyEcosystem Pilot 2: mySITE (data provision, visualisation tools, and ecosystem status indicators).....	124
Box 13-7: Standards, Data Models and Best Practices are useful for the success of each FAIR Principles: Findability, Accessibility, interoperability, and Reusability. Lessons learnt from mySITE	126

1 INTRODUCTION

Earth Observation (EO) involves a large range of data sources: not only from satellites, but also from ground-based and oceanographic observations collected through in situ, radars, and citizen sciences. From the angle of data science, EO includes data acquisition, data quality monitoring, data processing, science, statistics, and all types of computer sciences such as telecommunications, from data management to very big data management, cloud, web, mobile technologies, web analytics, computer graphics and visualization, and more. From the skills angle, EO includes soft skills such as human factors to face the complexity of the challenges it can support, the complexity of the Earth Observation products themselves, and of their production processes from data acquisition to product delivery. Human factors refer to environmental, organizational and workflow factors, human and individual characteristics which influence behaviour at work in a way that can affect health, safety and efficiency. To provide the right product at the right time with the right skills for the right decision, EO mobilizes multidisciplinary teams including experts from data technologies, sciences, computing, human factors, business, communication, and legal issues, who need to share a common understanding framework and a baseline of knowledge to cooperate efficiently.

This guide does not attempt to develop expertise on all topics but rather to share a holistic approach and understanding, to support the EO stakeholders in their global process from data to product delivery and in their interactions with their diverse background colleagues. Its goal is to mainstream a baseline of knowledge to support the community, intensify the connections between the different expertise and skills, and speed up the development process from the concept to a prototype and operations in a domain where the data constantly evolves and is very expensive to produce and process. The report might help specify some of the new or evolving terminologies and buzzwords resulting from the quick dynamics of the Earth Observation domain.

The report links to many external technical references to keep the holistic vision and understanding more readable, and to avoid quick obsolescence in a domain that is constantly and quickly evolving in terms of data volumes and data types, technological capacities, challenges, and impacts addressed, industrial networks and market structure, users' maturity. Information that can face quicker obsolescence is gathered in annexes and referred to in a more generic or conceptual way into the body of the document.

Based on the initial assessment of the e-shape pilots, the project has gathered an extensive understanding of the Pilot's teams' expectations and needs on one hand, and of the European Earth Observation resources they were using or planning to use on another hand (Figure 2). Considering the number of participants and their diversity in terms of nationality, scientific or technical background, gender, and age, e-shape partners can be seen as a representative sample of the Earth Observation community and the issues or expectations that they have expressed can be used as a catalogue of topics to address into the best practices. As a matter of fact, they cover all the workflow from data to product or service.

The development of an Earth Observation application requires addressing the 3 canonical scenarios:

1. how the users discover, access the data, or run the pilot
2. the new or improved EO service scope and development needs to clarify the interactions with the EO resources (platforms and data) used as an external resource
3. the publication, dissemination ... of the results of the new or improved service.

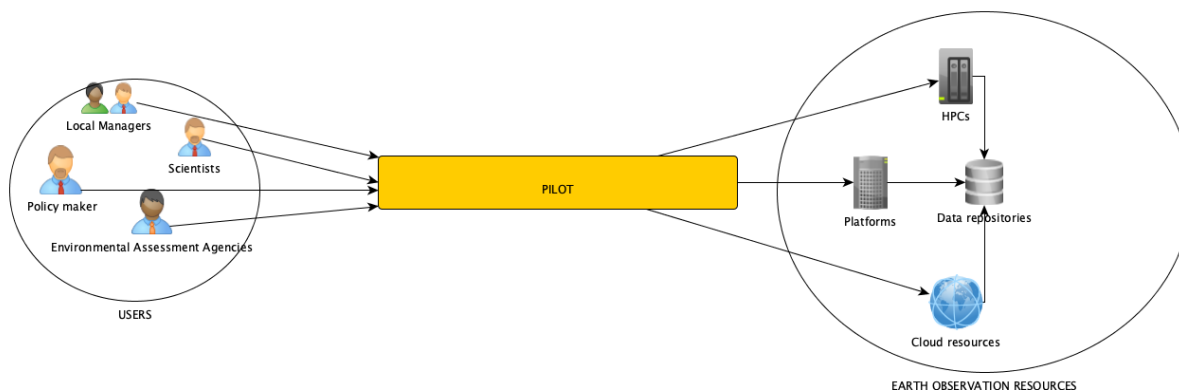


Figure 2: How do the user discover, access or run the pilot and how does the pilot interact with EO resources?

During your reading, access e-shape's Pilots: <https://tinyurl.com/e-shape-Pilots>

Other access to e-shape Pilot's 'knowledge granule':

- <https://e-shape.eu/index.php/all-pilots>
- <https://tinyurl.com/GEO-Knowledge-Hub>
- <https://tinyurl.com/GEO-DAB>

2 CO-DESIGN

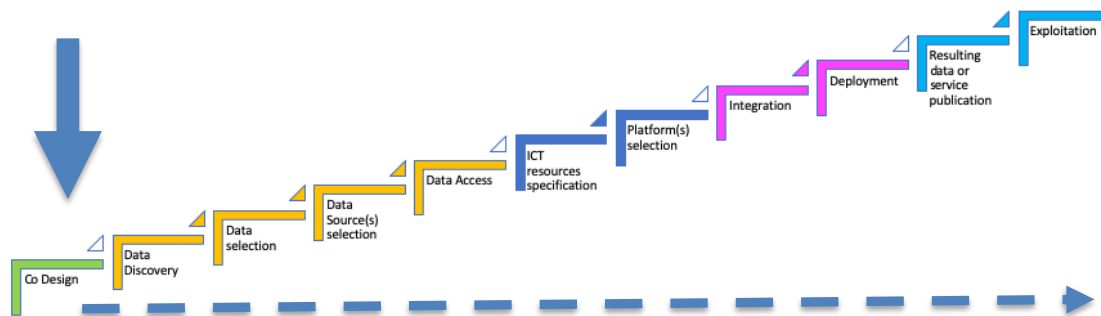


Figure 3: Co Design activities in the Development workflow

2.1 Introduction

Initially produced mainly for scientific goals, EO data are now made available to every economic actor, through ‘open-data’ policies. Socio-economic applications of this data seem to be diverse and promising for a large variety of socio-economic stakeholders: research communities, but also public authorities, private companies, academia, and citizens. However, in practice, developing usages from EO data seems to be particularly challenging. Perceived as highly technical, EO data, and to a certain extent derived services, remain largely underutilized despite their accuracy and therefore capacity to enable decision-making. Indeed, this effort could be schematically described as connecting various and highly heterogeneous socio-economic ecosystems: the ecosystem of EO data and the various ecosystems of potential usages, that do not share the same dynamics, time horizons (e.g. very long cycles to develop new measuring instruments compared to short timeline of actions in the data usage context), performance logics and competencies (e.g. data processing might require very specific technical expertise while data usages might also require specific domain expertise).

Co-design precisely aims at connecting these various and heterogeneous ecosystems of data and usages, through the development of EO-based services, supporting their dynamics in a long-term perspective, ultimately, ensuring the design and delivery of products meet user needs and skills.

In e-shape, a co-design model considering EO specificities is progressively designed and tested with e-shape pilots, through a dedicated work package (WP2). A first analytical framework has been built and described in D2.1, D2.2, and D2.3 deliverables especially highlighting that a co-design model adapted to EO context should involve two distinct phases:

- (1) a critical “diagnosis process” to identify the co-design needs, classified in four main types of co-design,
- (2) the implementation of co-design actions to address these co-design needs. e-shape has been able to build and test all 4 types of co-design actions with several pilots. This process is well described in D2.4, D2.5, 2.6, and D2.7 deliverables. All deliverables are accessible at the following link: <https://e-shape.eu/index.php/resources>.

In this section, we will describe the best practices that we have drawn from this work so that they can benefit the entire EO community.

2.1.1 Example of a diagnosis process to help the pilots to better structure their co-design strategy

Based on the analysis of e-shape pilots, a certain variety of co-design needs could be identified, leading us to define four main types of co-design:

Table 1 : Classification of co-design needs to grow an ecosystem of efficient service

	Overall context	Initial state	Blocking point to be addressed	Expected outcomes
Type 1	Adjustment between user and service designer	(a) Usefulness already identified on a first basis but to be enhanced. Usability to be enhanced. (b) Relationship with the user to be precisely defined but at least user willing to devote time settling it.	Establishing adapted relationships with specific users for <i>usefulness & usability assessment and enhancing</i>	(a) Expanded range of lists of requirements ensuring usefulness and usability (b) Cooperation modalities with these specific users clearly formalized
Type 2	Exploration for usage initiation	(a) Usefulness not well-known and/or (b) Relationship with the user appearing to be difficult to establish (uncommitted users)	Establishing adapted interactions with user communities for <i>usefulness identification</i>	(a) Expanded usefulness of the service (b) Expanded list of relevant stakeholders to interact with
Type 3	Engineering for service operationalization	(a) Requirements for usefulness and usability established. (b) Relationships with some users established.	Establishing adapted relationships with relevant partners for <i>extensive usefulness & usability realization and operationalization of the service</i>	(a) Clarification of the service structure (parts ready to be operationalized, parts needing further exploration) (b) Cooperation modalities between R&D and operationalization entities clearly formalized
Type 4	Exploration for usage expansion	(a) Existing service (usefulness & usability established for at least one use case) (b) Relationships already established with existing users.	Establishing adapted relationships with existing & potential new users for <i>usefulness reinvention</i>	(a) Expanded range of potential alternatives for future usages (which usefulness for which actors) (b) Cooperation modalities and supports for interactions (proofs-of-concept) defined for existing and new users

To carry out the co-design needs analysis meetings, e-shape has developed a grid for analyzing the pilots, representing each of them in a so-called ‘data-information-usage’ framework (see Figure 2). The objective of this analysis is to draw up an overall picture of the pilot context and identify the possible blocking points hindering the development of new EO usages. This especially includes the analysis of the data transformation processes (from data sources to a generic form of information that can be used in multiple usage contexts, up to value-embedding usages) and the various stakeholders involved all along this data-information-usage chain. Five main points of analysis have been specially considered and are represented in the figure below.

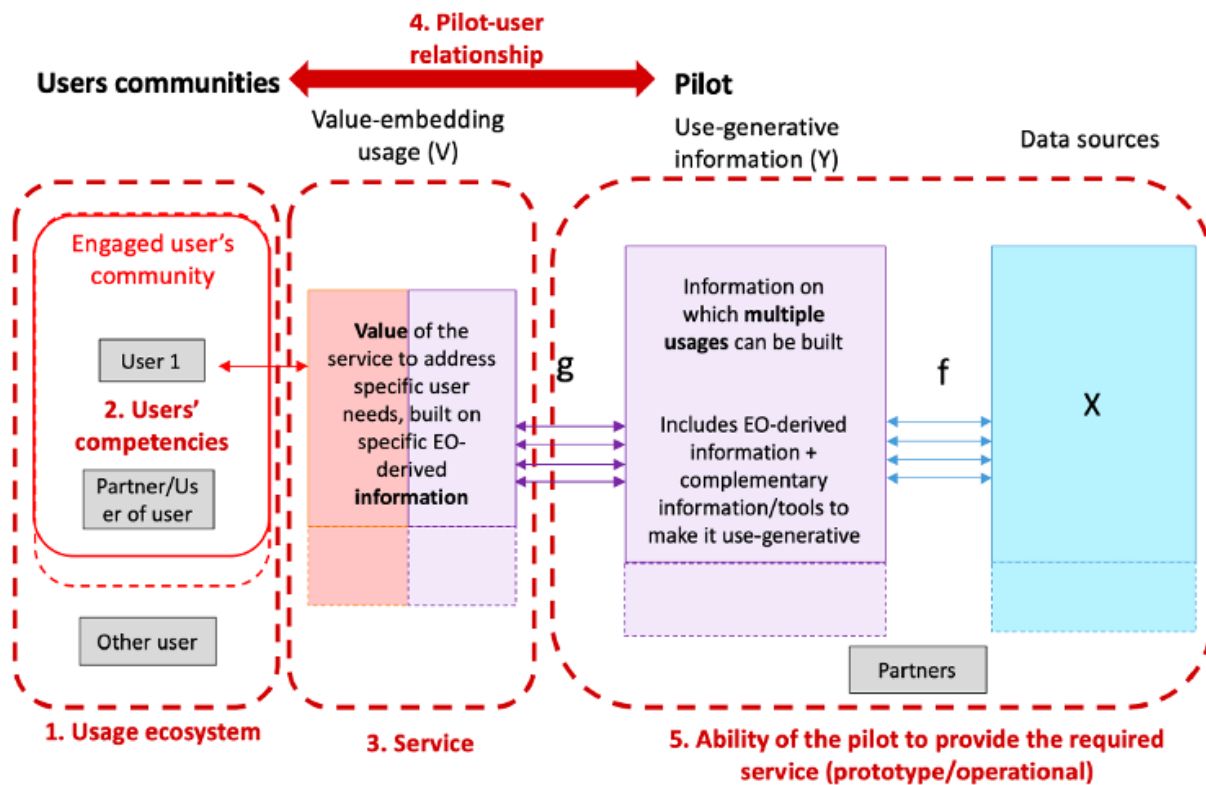


Figure 4: Representation of the "data journey" for the targeted state based on the data-information-usage framework

The 37 pilots of e-shape benefited from this diagnosis phase. This has proved to be particularly helpful for pilots to better structure their development approach, identifying which types of co-design actions would be relevant at which time horizon.

A first version of a self-diagnosis tool has been developed by WP2 to help e-shape pilots to carry out their own co-design needs diagnosis, consisting of an excel sheet and guiding documents. The tool was used by the 5 last onboarded pilots only as it was not ready in the beginning of the project and the 5 last onboarded pilots did not benefit from WP co-design mainly support due to a lack of time.

This tool has been built to allow all EO-based service developing projects to clarify their strategy by eliciting the status of their knowledge on the usage ecosystem and the considered development paths. This analysis also results in identifying possible blocking points calling for specific co-design needs.

The analysis of the pilot is done through a multitude of questions compartmentalized according to the 5 sections of the data-information-usage framework:

1. Usage ecosystem: user community, contact point, general context, and position in the user community
2. Users' competencies: category of user (e.g., EO expert, non-EO expert, software dev, etc.), user's daily use of data-based tools, additional support to users
3. Service: type of service, short description, EO-data derived information on which the service is based, maturity, level of access (e.g., restricted to the owner, open access, partners, etc.), lists of requirements, need of customization, interest of the user and service integration in user's operations

4. Pilot-user relationship: direct contact with the user, level of engagement, history of the relationship, expected inputs from the user, cooperation modalities, and feedback loops
5. Ability of the pilot to provide the required service (prototype/operational): role of the different partners involved in the development of the service, existence of a first functional service, upscaling challenges, dedicated operationalization team, cooperation modalities, and resources for operationalization

Depending on the answers given, the pilot can know which type(s) of co-design will benefit them. Here is a table showing the sections of the questionnaire that should be referred to, to identify the type(s) of co-design that the pilot needs:

Table 2: Distinction between 'quick-fit' and 'resilient-fit' perspectives for the 4 types of co-design

	User communities	User competencies	Service developed by the pilot	Pilot-user relationship	Ability of the pilot to provide the required service (prototype/operational)
Co-design type 1	X		X	X	
Co-design type 2			X	X	
Co-design type 3	X		X	X	X
Co-design type 4			X	X	X

Pilots that did not need WP2 support to conduct their co-design actions did benefit from WP2's upstream work and expertise:

- “We had a very good collaboration with WP2, and we have four pilots that also had bilateral meetings and discussions on how to develop co-design. Indeed, we have been doing more than initially hoped and this has helped to include and to co-design with some users that they (i.e., the pilots) found in the course of time”, Alexia Tsouni (NOA, SC6 coordinator alternate)
- “The user guide helped us to look at the service from another perspective, upside down, and technically the requirements were not clear but now they are” Annelies Hommerson (Water Insight, S5-P5 pilot alternate) See deliverable D2.6. for more details: (<https://e-shape.eu/index.php/resources>).

Here is an example of a data-information-usage framework completed thanks to the information provided by a pilot using the initial assessment questionnaire:

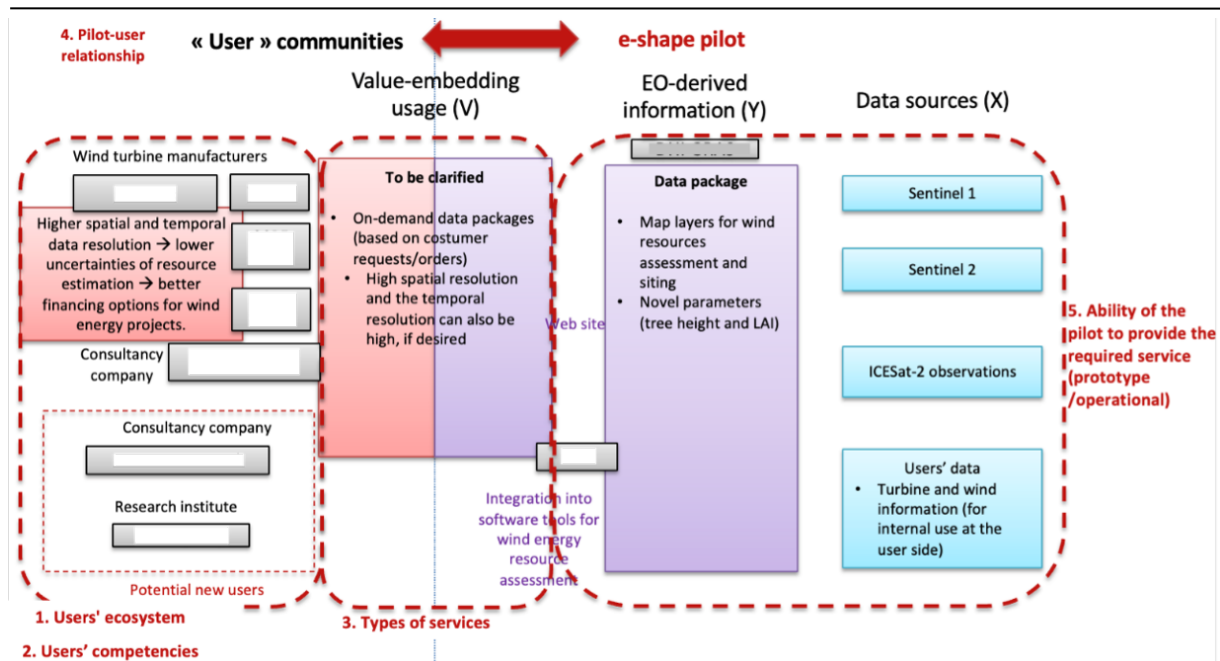




Figure 5: Anonymized data-information-usage framework completed using the answers to the initial assessment questionnaire

In e-shape's EO co-design method, each co-design type is supported by a series of workshops designed to progressively shape and consolidate 'building blocks' of the long-term development of the pilot's strategy, intertwined with the evolution of both EO and usage fields. Based on the experimentations carried out in e-shape, two dimensions appear as particularly critical for the success of co-design actions in a long-term perspective:

- Key insight 1: the co-design actions should not only focus on the design of the service but also on the design of the relationships, i.e. 'co-design' has to design the 'co'. The protocols of the workshops integrate this aspect by always organizing a final phase dedicated to building agreements for future cooperation between participants.
- Key insight 2: the co-design actions developed by WP2 aim at establishing a 'resilient fit' between participants, rather than a 'quick fit':
 - 'Quick-fit' actions would focus on finding one type of interaction between data and usages ecosystems (single list of requirements with one user, in a punctual relationship).
 - Whereas, 'resilient-fit' actions aim at generating a range of alternatives (regarding the lists of requirements, the stakeholders involved, the types of partnerships), allowing a better adaptation to future surprises or unexpected constraints.

The difference between these two types of actions can be illustrated by the metaphor of a plant that is more resilient as its roots' network is expanded, allowing the plant to adapt to various types of water conditions (see table below). This point appears to be crucial to foster the use of EO in a long-term perspective, as pilots will have to deal with constant evolutions of both the EO field and the different usage fields.

Table 3: Distinction between 'quick-fit' and 'resilient-fit' perspectives for the 4 types of co-design

	"Quick-fit" actions	"Resilient-fit" actions
General description	 <p>Focus on finding ONE type of interaction with the ecosystem (single list of requirements with one user, in a punctual relationship)</p> <p><i>If roots only at surface level: plant only grows if water is easily accessible</i></p>	 <p>Generating a range of alternatives (regarding the lists of requirements, the stakeholders involved, the types of partnerships) for a better adaptation to future surprises</p> <p><i>Expanded root network: plant more resistant to various water conditions</i></p>
Type 1	Finding ONE satisfying list of requirements with one specific user	In order to end up with a robust list of requirements, exploring a range of potential services at different time horizons and related cooperation modalities
Type 2	Finding ONE relevant user to interact with	Progressively building a better understanding of the usage ecosystem and cooperation agreements with a portfolio of relevant actors
Type 3	Building the engineering for the operationalization of one service	Building relationships with relevant partners to ensure a continuous investigation on modules to be operationalized/to be explored
Type 4	Merely asking existing users what they would dream of	Setting-up a joint program for long-term exploration of new usages with existing and new actors (identification of obstacles, research efforts to be made, 'stimulating' proofs-of-concept, etc.)

A specific protocol has been designed for each type of co-design action and has been experimented for all co-design types. WP2 was able to develop a user guide for each of the three first types of co-design actions. The user guide for co-design type 4 is under development. This allows pilots which did not ask for WP2's help to follow the WP's recommendations as well as possible to reach a "resilient-fit" or to have material to better manage their activities. Each user guide consists of a folder with 2 templates (workshop and formalization of outcomes) and a folder with the same documents but filled with information to show the example to follow. Templates encourage pilots to organize the workshops into phases that are themselves punctuated by a series of guiding questions

For further details on protocols please refer to D2.6 and D2.7 deliverables following this link: <https://e-shape.eu/index.php/resources>.

2.1.2 Real-world application and added value

A total of 5 co-design actions were conducted punctuated by various workshops:

- 1 co-design type 1 workshop with S2-P3 pilot ([Health Surveillance Air Quality](#) within the [Health Surveillance](#) Showcase) aiming at building an air quality & health surveillance platform for current and future operations of Athens' actors
- 3 co-design type 2 workshops for S3-P3 pilot (Offshore wind resources within the [Renewable Energy](#) showcase). Each one of them was done with a different stakeholder and aimed at leveraging knowledge & experience, exploring the range of usefulness of the pilot's service and related actors of the ecosystem
- 1 Co-design type 3 for S3-P2 pilot (High PV penetration at urban scale). It was conducted with the presence the pilot and its partner in charge of the engineering and commercialization of its services. The topic of the workshop was: Based on the concrete cases identified in the preliminary phase, clarifying the parts of the service to be operationalized/to be explored & the associated cooperation modalities between the service development team and the operationalization team

- 1 workshop conducted by the pilot without the support of WP2
- Co-design type 3 for S4-P2 ([mySITE](#)) and S4-P3 ([myVARIABLE](#)) pilots. This use case is very specific as it is composed of 3 pilots that develop modules or portals with the objective of being interconnected and forming the EcoSense platform. 1 workshop with both pilots aiming at clarifying the parts of the service to be operationalized/to be explored & the associated cooperation modalities
- Co-design type 4 for S5-P4 ([Sargassum detection for seasonal planning](#)). 3 preliminary sessions and 2 workshops were necessary to carry out this action
 - Workshop 1's objective was "Sharing CLS & CERMES knowledge" on the sargassum ecosystem to build a sustainable CLS - CERMES relationship and further stimulate the sargassum ecosystem.
 - Workshop 2's objective was "Exploring the business model of meteorological institutes to build a sustainable CLS - CERMES relationship & further stimulate the sargassum ecosystem"

Co-design actions have often brought a double benefit to pilots: designing the “co” (further information on the collaboration modalities) and a variety of development paths at different time scales. To properly set up and present these outcomes, WP2 has developed an adapted support:

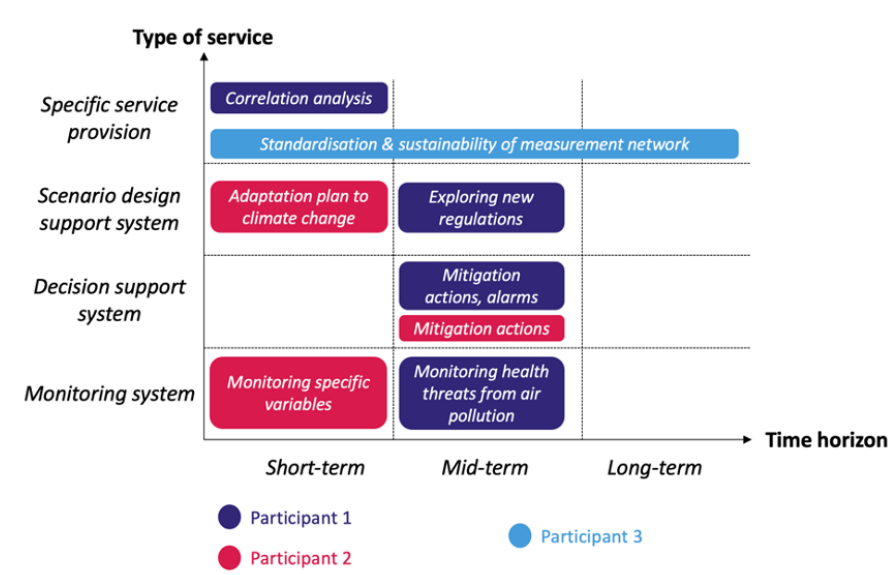


Figure 6: Graph synthesizing co-design type 1 outcomes in a 'resilient-fit' perspective.

For further details on protocols and the results on the co-design actions we have conducted please refer to D2.6 and D2.7 deliverables following this link: <https://e-shape.eu/index.php/resources>

To illustrate the added value that these co-design actions have had, we have selected some feedback we have received from pilots:

- “The workshop served as a means to formalize relationships and find synergies between workflows and users, propelling us to officially pursue partnering with National Public Health Organization and the Ministry of Energy and the Environment to discuss and share data, and contribute to the development of a national health observatory”, Evangelos Gerasopoulos (NOA, S2-P3 pilot leader)
- “I’m really satisfied and impressed with the support regard Raphaëlle and her team [...], especially I wanted to highlight the treatment and the analysis of the outcomes of the different workshops we

had. It was very helpful to have some tools and diagrams like we've seen today to organize all the inputs [...] For me, it was really eye-opening that we could use it in such a broad way to look at all sorts of possibilities rather than trying narrow down what we wanted to do. It was more about broadening out and gathering lots of ideas and inputs.”, Merete Badger (DTU Wind Energy, S3-P3 pilot leader)

- *“This exercise has proved to be useful as in 3h we have structured our working plan for the next 6 months in a clear way.”* (Etienne Wey, Transvalor, member of S3-P2 pilot) *“We learned a lot definitely. It's something which dealt with some tremendous unknown things that we learned by talking to you [i.e., WP2 team] through this process”*, Lionel Ménard (O.I.E., member of S3-P2 pilot)
- *“We learned a lot with the support of WP2 and the experience gained in co-design action will be for sure reinvested in the future developments probably by targeting other types of users [...] but willing to strengthen their position”*, Marion Sutton (CLS, S5-P4 pilot leader)

2.1.3 Self-diagnostic tool references

The self-diagnosis excel file can be downloaded following this link: https://e-shape.eu/images/co-design/Initialassessment_questionnaire.xlsx

2.2 On-going reflection on further co-design routinization

From our first observations, it appears that the self-diagnosis tool is helpful to initiate the diagnosis process, but a telco with WP2 was still required to finalize the analysis. Further use and work are needed to improve this self-diagnosis tool to allow pilots to carry out their diagnosis in total autonomy.

e-shape's pilots' cases are heterogeneous and a pilot needing a type 3 co-design will not answer in the same way as a pilot needing another type of co-design. Moreover, each pilot has its own environment and modelling, having specific answers as an example is not a viable solution for an uptake on a large scale and without the support of a dedicated team. Thus, it seems essential to improve the knowledge of co-design among all the actors of the sector. This could be achieved through training provided by a team dedicated to co-design at one of the reference institutions of Earth observation such as EuroGEOSS or ESA.

2.2.1 Co-design routinization beyond e-shape

Identified paths for the future dissemination and routinization of the specific co-design framework for EO developed within e-shape are:

- Guidebooks (diagnostic tool & guidelines for workshops)
- Developing co-design as-a-service (e.g., training of consultancy companies)
- Establishing co-design as a critical component of EuroGEO/GEO, e.g.:
 - Diffusion of best practices
 - Setting-up co-design training for the EO community
 - Ensuring co-design quality (labelling system)
 - Funding future research on co-design advances

Achieving a certain level of standardization of the method used to co-design in the Earth observation sector is an important lever for dissemination. This aspect was detailed in D2.8 and D2.9 deliverables.

Box 2-1: Co-design. Lessons learnt from Showcase 5: [Water resources management](#) Pilot 6 [EO based phytoplankton biomass for WFD reporting](#) (WI)

The Earth Observation based phytoplankton biomass for WFD reporting pilot has worked according to the co-design methods for one of the demonstration sites, using most of the guidebooks. Although we and our user had an idea of what we wanted to implement, we organised a session with other stakeholders, to take them along in the development process. Alpha products (ideas, mock-ups) were used as a starting point for the discussions. They came up with good questions and additional ideas, which led to the next development phase. To present the results ('beta products'), a second session with the involved stakeholders was organised. Results were discussed and led to statements on further use of Earth Observation based mapping after the project end. This indicated that the co-design methods generated not only interest for the pilot service, but also a broader base of support for Earth Observation based methods. This was a very favourable outcome, and we will continue using co-design methods.

Lessons learnt

- it is very useful to take not only the contact point, but also other stakeholders along in the development process
- even though you think you already know what is wanted, better start with mock ups and examples and improve gradually, it might save time, but maybe more importantly it makes the users feel 'co-owners' of the result

References on co design:

- Le Masson P, Weil B, Hatchuel A (2017) Design Theory - Methods and Organization for Innovation. Springer Nature.
- <https://e-shape.eu/index.php/resources#sppb-modal-1679641282079>

3 DATA, PROCESS, AND APPLICATIONS DISCOVERY

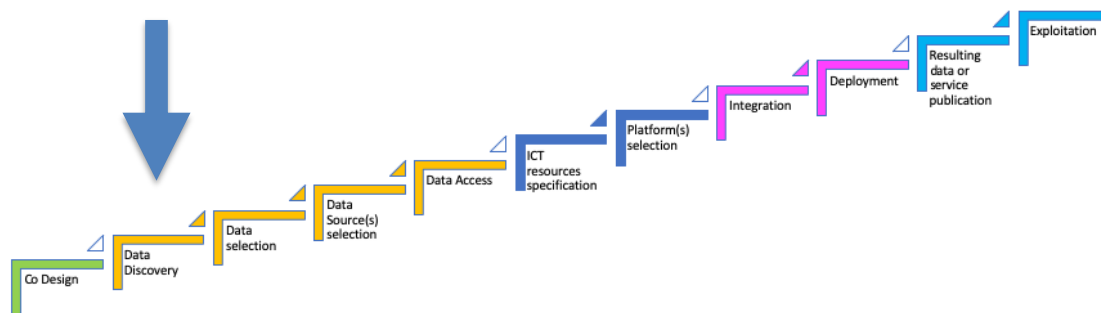


Figure 7: Data Discovery in the Development workflow

3.1 Introduction

The first critical step in any data analysis or development is to acquire enough reliable data. This involves identifying, exploring, and understanding data sources, including their structure, content, and quality. This process is known as data discovery and is essential in any data management or analytics project. Data discovery enables data professionals to identify data sources, assess data quality, and understand the relationships between data elements. It ensures that data is accurate, complete, and relevant to the organization's needs.

Data discovery techniques may include data profiling, data lineage analysis, and data cataloguing. Data profiling involves analysing the data to understand its structure, completeness, and consistency. Data lineage analysis involves tracking the flow of data across systems and processes to understand where it comes from and how it is used. Data cataloguing involves creating a searchable inventory of data assets that enables data professionals to easily locate and access the data they need.

To reduce search time and improve the final outputs, improving data discoverability is crucial. Standardized metadata and centralizing libraries can achieve this. Locating the relevant data, making sense of it, and evaluating whether it is trustworthy remains a challenge. Metadata, which are data about data, are vital at this stage of the filtering process. They describe the properties of a dataset and can cover various types of information. Metadata facilitate file discovery and cataloguing, ensuring that datasets are easily discoverable within a database or online. Metadata catalogues use metadata standards to ensure that they are useful to different users and are "machine readable."

From Data Portals to Marketplaces, Catalogues have improved significantly in recent years, taking advantage of metadata standards. They can target expert communities in specific domains or be generalists to any type of topic. They can specialize in a specific type of data or manage any type of Earth Observation (EO) data, such as in-situ, Citizen Science, Remote sensing, radar, etc. They can focus on a specific region or be global. Catalogues can target scientific research communities, offer a demanding service level agreement for disasters and businesses, or manage different levels of authentication and authorization. The catalogues can be part of a platform offering Cloud capacities in addition to the data catalogue, to process the data locally. The same data can be discoverable via different catalogues. This redundancy is an opportunity for the user searching for data - who will find it more easily, and for the data provider - who will have more opportunities to grow their user community, reaching even an unexpected audience. The remaining issue for the data provider is to select the most relevant catalogues. Standards are available to facilitate and minimize the workload for data publication

in several catalogues. This will be discussed later in the "Resulting data or service publication or dissemination" chapter.

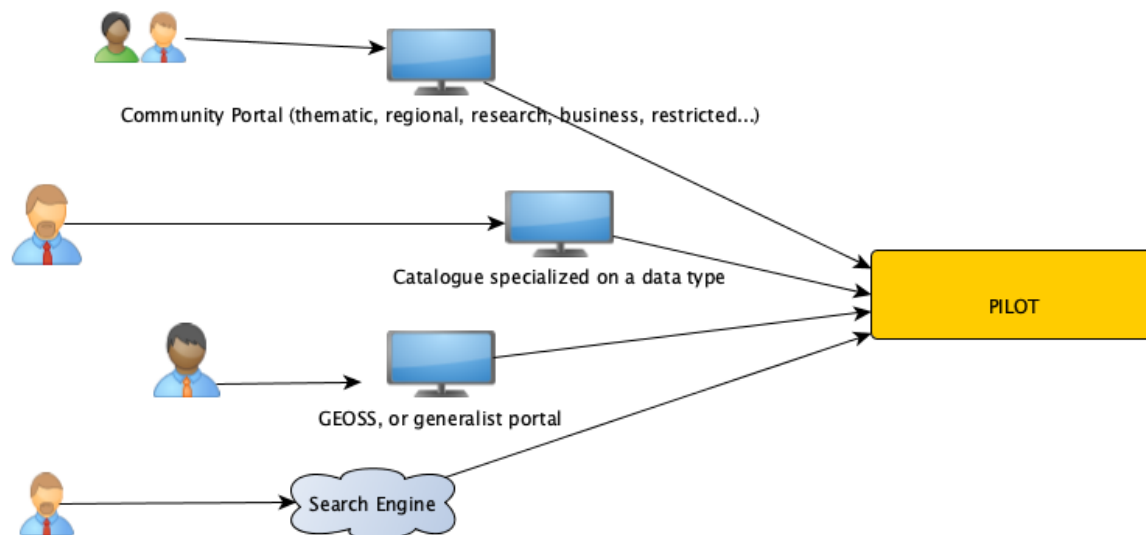


Figure 8:User-Pilot interaction

Discovering open remote sensing data is usually easy. The issue will mainly be selecting the data source that brings the temporal depth, update frequency, access and processing capacities, and quality of service that is needed. In-situ discovery is more challenging because there is a lack of global networks, the data providers are far more fragmented, and a lot of data is just not shared. Data Spaces could address this issue in Europe in the future but at the time being, they are not available. In a world where many issues are global, where help in case of crisis can come from remote locations, the Earth Observation community is urging in-situ data providers to make their data more discoverable and facilitate access via more global collections. In-situ is needed to calibrate and validate remote sensing products, initialize, validate, and calibrate numerical models, train machine learning algorithms, and assess the quality of the value-added products or processing. It is critical to the Earth Observation industry and to all societal challenges impacted by Earth Observation.

In e-shape, the [Agriculture](#) and the [MyEcosystem](#) Showcases have addressed this issue via major contributions to Agrostac and DEIMS-SDR in-situ databases respectively for Agriculture and for Biodiversity.

The effort to build in-situ global catalogues is currently pushed by the domain communities as building a global in-situ catalogue for all domains is still an inaccessible grail.

The amount of Earth Observation Data providers increases every day. Today any citizen, any object, can become a data source contributing to Citizen Sciences data, to the Internet of Things or to both. In the first step, each of these data production processes often results in a data silo related to the initial team, project, or initiative. As a data provider, publishing in sustainable portals and repositories can bring, with minimal effort, the opportunity to make the data used and reused by a growing audience, bringing sustainable benefits to the data producer and to the global community.

As a data user, if your application only uses one source of data it might not be a problem but most Earth Observation applications require different types of data that will quickly span several systems. Fragmented and incompatible data or data encoded differently slow down innovation.

As the current Earth Observation scenarios are more and more sophisticated, requiring more heterogeneous data from heterogeneous communities using different standards, federations of catalogues or various types of Data management strategies such as data warehouses, data lakes, data spaces, data hubs, and/or data brokers are used to store, organize, and analyse data. However, they differ in their operation, purpose, and practical use and interoperability remains key at all stages: discovery, access, processing, and dissemination.

Data warehouses are the older data structures that provide discovery and access to data that is collected and structured to address specific scenarios. Data lakes are flexible storages for any type of raw data, structured or not, without specific data organization. It requires good data governance to avoid becoming a "data swamp". Data spaces are built for a specific community, have a strong focus on data sovereignty, trust, and interoperability and Data hubs connect different systems and allow data discovery and access from heterogeneous catalogues and data infrastructures.

Data lakes, data spaces and data hubs facilitate innovation, data access, and data reuse. Data hubs often provide access to Domains or scenarios oriented spaces to develop a user experience that looks like the Data warehouses. Data lakes can be considered more flexible to add new data, Data hubs are more flexible for adding new business scenarios. The data hub harmonizes the metadata from different data silos, then thanks to an indexing strategy will facilitate and accelerate data access. It requires fewer data skills from the user than a data lake that is more for data experts. Data brokers are data aggregators. Gartner defines it as: "A Data Broker is a business that aggregates information from a variety of sources; processes it to enrich, cleanse or analyse it; and licenses it to other organizations. Data brokers can also license another company's data directly, or process another organization's data to provide them with enhanced results. Data is typically accessed via an application programming interface (API) and frequently involves subscription-type contracts. Data typically is not "sold" (i.e., its ownership transferred), but rather it is licensed for particular or limited uses. (A data broker is also sometimes known as an information broker, syndicated data broker, or information product company.)" (source: <https://www.gartner.com/en/information-technology/glossary/data-broker>)

For more than 10 year, the Group on Earth Observations (GEO) supports discovery via the GEO System of Systems (GEOSS) portal providing access to a very large number of Earth Observation datasets globally. Since September 2018, Google Data Search also addresses this issue of locating data freely available for use. The service is targeted at scientists and data journalists.

Box 3-1: Data discovery. Lessons learnt from the [mySITE](#) Pilot [MyEcosystem](#) showcase

The pilot [mySITE](#) (data provision, visualisation tools and ecosystem status indicators) of the [MyEcosystem](#) showcase aims to mobilise and valorise in-situ data collections from long-term observation facilities mainly managed and operated by organisations linked to the eLTER network. Understanding the context of observations as well as getting information on the observation facilities supports e.g. to the EO community. This would be valid for planning targeted in-situ campaigns or retrieving information on data collected at these sites. The DEIMS-SDR (Dynamic Ecological Information System - Site and Dataset Registry, [Wohner et al., 2022](#)) is a web catalogue to document and share information on long-term observation facilities with currently approx. 1200 sites described on a global scale. DEIMS-SDR provides a wealth of information for a wide range of sites, including each site's location, ecosystems, facilities, parameters observed, and research themes. It also links and provides access to a growing number of datasets and data products associated with the sites. All sites and dataset records can be referenced using unique identifiers, the [deims.id](#), that are automatically generated and can be used across different catalogues to unambiguously identify co-located sites. Information on sites can be found via keywords, predefined filters or as a map search.

This pilot implementation has emphasized the following lessons learned:

1. information on the observational context of managed long-term observation sites including contact details is benefiting different stakeholders, including the EO community,
2. unambiguous and persistent identification is needed to allow for referencing across different in-situ networks and research infrastructures in the terrestrial and transitional-waters domain,
3. providing high-level documentation on observation capabilities fosters data mobilisation and enhances data findability

References on the DEIMS-SDR

- Wohnner, C., Peterseil, J., Klug, H. 2022. Designing and implementing a data model for describing environmental monitoring and research sites. In *Ecological Informatics*, 70, p. 101708). Elsevier BV. <https://doi.org/10.1016/j.ecoinf.2022.101708>
- Deims-SDR portal: <https://deims.org/>

3.2 From data download then process to data processing via online applications then download

Data volumes are growing exponentially. In 2019, ECMWF estimated that 10 years later, they would get 10 times more observational data per day, 2000 times more model data per time step, 25 times more forecast product data per day in the critical path, 100 times more data archived per day, and 30 times more data sent to customers per day in the critical path². Eumetsat announces in the future an increase of 50 times more data with the next satellite generation. With this exponential growth of the data, data downloading can be too slow or challenging and the growing new paradigm is to upload applications and run them close to the physical location of the data. This growth of data also requires new investments in technology, science, and research.

² source: https://www.ecmwf.int/sites/default/files/medialibrary/2019-01/02_DELLACQUA_INDUSTRIYDAY_20190116.f.pdf

Two capabilities have to be considered: discovering data, downloading it, and running local processing or discovering applications, uploading them near the data location, and processing near the data. If the processing is efficient, this allows moving from a concept of pre-processed data to data processed on the fly with a customized application, only when needed which is more and more relevant when the data is updated very frequently. In this case, the catalogues have to evolve from Data catalogues to Applications catalogues and from data access to processing near the data.

We are currently in a transition phase where the two approaches co-exist: Download the data then process or push the application in the cloud to process near the data. The Cloud was first used as Data storage to make the data more accessible, it is also used now as a processing resource to make the data more usable.

Box 3-2 : Data download to processing. Lessons learnt from [myVARIABLE](#) Pilot [MyEcosystem](#) showcase

The need to move the application close to the data is a crucial consideration in modern Earth observation applications, due to the ever-increasing volumes of data needed to power them. In the case of the [myVARIABLE](#) pilot project, the algorithms used to calculate Essential Biodiversity Variables require data from remote sensors ranging in size from tens to hundreds of GB in the case of high-resolution time series, which must be accessible to the algorithms locally. This poses a challenge when structuring the data in multidimensional array format and packaging the service in a Docker container, as the resulting file is too large to be handled and downloaded efficiently.

To mitigate this challenge, we propose several solutions, including data compression techniques, data partitioning, and the use of edge computing architectures. Compressing the data can reduce its size and make it easier to package, while partitioning can enable faster and more efficient access. Edge computing architectures, which involve placing computing resources closer to the data source, can also provide a solution by enabling faster processing and reducing on-demand data transfer and analysis times. However, each solution brings its own challenges, such as increased complexity, reduced reliability, or higher cost. Therefore, careful consideration must be given to the specific requirements of the Earth observation application, based on a market study of the users of our service, including the different Biodiversity Observation networks such as GEO BON, EuropaBON, and TAO, among others.

Box 3-3: Data download to processing. Lessons learnt from Showcase 4 [Water resources management](#) ,Pilot 5: [Sargassum detection for seasonal planning](#)

During the first phase of the e-shape project, the objective of the [Sargassum](#) pilot was to move an existing operational chain to the cloud, in order to benefit from higher computer resources and accessibility of the data. Indeed seasonal planning implies the processing of long time series of sargassum detection (one year minimum) and over a large area (the entire Tropical Atlantic), which represent a large amount of satellite data to acquire and process (10 Terra octets of OLCI Sentinel-3 data to produce 3 Terra octet for one year).

For our pilot, the selection of the DIAS was done in 2020 according to two main criteria:

- the availability of a Kubernetes cluster with access to registry docker
- the access to archive satellite data from 2019.

While the prices proposed were similar, the technical offer was different, probably also in the early stage of the DIAS offer construction. However, the access to the archive data was a problem in all of them, and we had to acquire the Sentinel-3 data through an external source at NASA.

3.3 Community Portal, GEOSS, Google Data Search

e-shape has encouraged the e-shape pilots to publish their products in different portals to upscale their audience. Most of the pilots if not all have published their results in Community Portals, in the GEOSS portal, and the GEO Knowledge Hub and some are discoverable via Google Data Search. All the pilots are also discoverable via EoMall, EoWiki and as INSPIRE-compliant metadata records in the webservice-energy catalogue, harvested by the GEO DAB and the GEO Knowledge Hub. Of course, they are all documented in the e-shape project portal under the form of “id cards”. Web analytic tools can then be used to analyse the efficiency and effective engagement of the audience via the different channels.

Box 3-4: Community portals et al. Lessons learnt from Pilot [Forestry conditions](#) (more efficient forestry operations with lower environmental impact and carbon emissions) from the [Climate](#) showcase

The [Forestry conditions](#) (more efficient forestry operations with lower environmental impact and carbon emissions) pilot has developed a service supporting the forestry sector (harvesterseasons.com) and has disseminated the service promotion via several platforms and portals. The service has been using Google Analytics in the background to analyse the impact of all dissemination activities respectively. It is important to state that the analytics service requires a statement of privacy policy, about what user information is collected for statistical analysis. In the case of this e-shape pilot, it is only basic information on location, IP, and acquisition channel. But even with only basic information Google Analytics provides a good understanding of users, service usage, and the impact of dissemination activities.

Throughout the project timeline, it could be clearly seen that each promotion activity has impacted a rise in service usage. Most importantly the service is findable directly and prominently via Google search. Additionally, to just name a few promotion activities, the [Forestry conditions](#) pilot had the [Climate](#) showcase webinar series, individual service webinar, participation at users events for the Finnish forestry sector, and presentation at conferences. Each of these activities helped to raise awareness of the forestry service and was directly visible in click numbers.

Through Google Analytics it was also possible to investigate various user acquisition channels. The [Forestry conditions](#) pilot has been promoted through various channels. Harvester Seasons' own LinkedIn and YouTube profile, providing monthly service updates and information, helped most effectively to bring users to the service platform. Articles at the portals of the Finnish Meteorological Institute, Copernicus, E-shape, as well as WEkEO DIAS (Harvester Seasons is one of the WEkEO use cases), are well clearly visible as strong user acquisition channels according to Google Analytics. Last but not least the [Forestry conditions](#) pilot is findable via GEOSS as well as the national Finnish GEOdata portal <https://kartta.paikkatietoikkuna.fi/>. Even though those platforms are not yet well known in the forestry community it could be shown via Google Analytics that the Harvester Seasons service acquainted at least some visits via those platforms.

As a summary, a major lesson learned from this pilot, is that dissemination activities have to clearly include various channels to promote a service and not only one. Using a conglomerate of all kinds of actions for promotion like GEO portals, Google, Networking platforms, and community events is very important for raising awareness throughout different user groups, upscaling efficiently the outreach, discoverability, and use of the data. It can open unexpected markets. The analytics tools enable the analysis of most efficient channels and events to reach the biggest audience revealing the real value of the data and way to optimize its marketing.

Each portal addresses different publics and can require some different flavours of metadata. As it is critical to the success of the upscale strategy the topic is detailed in a specific chapter later in the document "Resulting in data or service publication or dissemination".

GEOSS, NextGEOSS and GEO Cradle are data portals that have been designed for the GEO community.

Box 3-5: Community portals et al. Lessons learnt from the [mySITE](#) (data provision, visualisation tools, and ecosystem status indicators) Pilot [MyEcosystem](#) showcase

A further possible contribution to the GEO Knowledge Hub would be the site catalogue developed by the pilot using DEIMS-SDR and its extension in the e-shape project. The work done contributes to building a global catalogue of terrestrial observation sites as addressed by the GEO community. Information collected on the long-term observation facilities could be shared with the GEO Knowledge Hub. Currently, the [mySite](#) pilot with its tools DEIMS-SDR and EcoSense is shared. Further options to share the data set (published as REST-API as well as OGC services) need to be explored and would be work beyond e-shape.

3.4 Standards for Data Discovery³

3.4.1 Opensearch

OpenSearch is "a collection of simple formats that allow publishing of search results in a format suitable for syndication and aggregation and the sharing of search results. It is a way for websites and search engines to publish search results in a standard and accessible format. The OpenSearch description format allows the use of extensions that allow search engines to request a specific and contextual query parameter from search clients. The OpenSearch response elements can be used to extend existing syndication formats, such as RSS and Atom, with the extra metadata needed to return search results.

OpenSearch is intended as an M2M (machine to machine) interface that provides a standardized way of 1) submitting search queries and 2) receiving results. The results are returned as XML.

The OGC OpenSearch Geo and Time Extensions specify the Geo and Time extensions to the OpenSearch query protocol. The OpenSearch description document format can be used to describe a search engine so that it can be used by search client applications. Services that support the OpenSearch Specification and the Geo and Time extensions defined in this document are called OpenSearch GeoTemporal Services. <http://www.opengis.net/def/docs/10-032r8>

References on Opensearch

- Gonçalves, P. 2016. OGC® OpenSearch Extension for Earth Observation. 13-026r8. <http://docs.opengeospatial.org/is/13-026r8/13-026r8.html>
- Gonçalves, P. 2010 OGC® OGC: OGC 10-032r8, OGC® OpenSearch Geo and Time Extensions: <http://docs.opengeospatial.org/per/19-020r1.html#OGC10-032r8>

3.4.2 Schema.org

[Schema.org](#) is a collaborative, community activity and a reference website that publishes documentation and guidelines for using structured data mark-up on web pages (called microdata) with a mission to create, maintain, and promote schemas for structured data and standardize HTML tags to be used by webmasters for creating rich results (displayed as visual data or infographic tables on search engine results) about a certain topic of interest.^[2] It is a part of the semantic web project, which aims to

³ In this section please click the underlined text to access expert-level details.

make document mark-up codes more readable and meaningful to both humans and machines. (source: <https://en.wikipedia.org/wiki/Schema.org>)

Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata, and JSON-LD. These vocabularies cover entities, relationships between entities, and actions, and can easily be extended through a well-documented extension model. Over 10 million sites use Schema.org to mark up their web pages and email messages. Many applications from Google, Microsoft, Pinterest, Yandex, and others already use these vocabularies to power rich, extensible experiences.

Founded by Google, Microsoft, Yahoo, and Yandex, Schema.org vocabularies are developed by an open community process, using the public-schemaorg@w3.org mailing list and through GitHub.

References on Schema.org

- Website: [Schema.org](https://schema.org)

3.4.3 The OGC Definition Server

The OGC Definitions Server is a Web-accessible source of information about things ("Concepts") the OGC defines or that communities ask the OGC to host on their behalf. It applies FAIR principles to the key concepts that underpin interoperability in systems using OGC specifications. The Definitions Server can be accessed at <https://www.opengis.net/def>

These things can be anything that is important in the course of interoperability around spatial information where the OGC plays a role in facilitating common understanding - either through publishing specifications or assisting communities to share related concepts. OGC uses stable web addresses (URIs) to unambiguously identify concepts in its specifications. The Definitions Server makes those URIs "work" - i.e. makes them dereference to a definition that can be used.

References on the OGC Definition Server

The OGC Definitions Server: <https://www.ogc.org/def-server#:~:text=The%20OGC%20Definitions%20Server%20is,in%20systems%20using%20OGC%20specifications>

4 DATA SELECTION

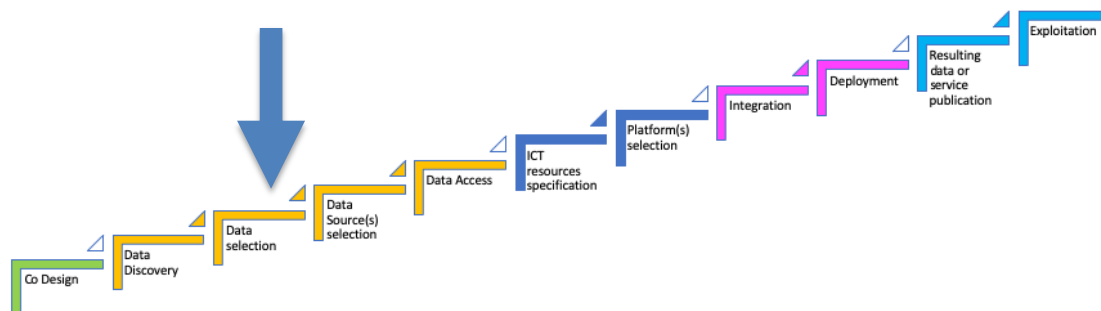


Figure 9: Data Selection in the Development workflow

4.1 Introduction

The major impulse to the EO industry has been the decision to open the Landsat Data in the 1970s followed by the Copernicus programme of Europe in the 2000s and other institutional programmes. Copernicus not only opens satellite data such as the Sentinel data but also offers Services organized in user-centric Thematic collections. The data is openly available for reuse and the only legal requirement is to credit the European Commission.

ESA sentinel online space (<https://sentinel.esa.int/web/sentinel/home>) provides details on the Sentinel data: Sentinel Missions, User Guides, Technical Guides, Thematic Areas, Data and Products Access, and Toolboxes.

Several data sets produced with different sensors onboarded on different satellites can be comparable or the same dataset can be accessible in different resolutions in space and time. Very often these different resolutions are attached to different access rights: low resolution can be accessible as open data and high resolution as paying data. When datasets are comparable but not provided by the same sensor or with different processes, a full process of data preparation including calibration and validation has to be implemented to be able to reuse it into a defined application producing comparable results. The fitness for use of the datasets will be characterized.

4.2 Data assets analysis

Some e-shape partners have raised the issue of the Data assets analysis in the initial assessment.

Analysing the assets of data is multi-dimensional. It can be related to its variety to mitigate the limitations of each EO data type, its volume (ex: Remote sensing data), a specific value it brings to a community (ex: a single observation in the middle of the oceans), its pre-processing done by the data provider to hide part of the complexity of the measurement process (Analysis Ready Data - ARD), its relevance for a specific Domain (EVs), its operational reliable delivery in a user centric way (ex: Copernicus), its coverage, its time frequency, its density or any of its characteristics and last but not least: its frequency of use.

The following paragraphs will raise awareness on part of these dimensions.

4.2.1 Different types of Earth Observation data

Earth Observation refers to many different types of data that have all assets and weaknesses. The good news is that they are fully consistent by essence as they all capture some measurement of the same reality, and they are complementary as their strengths and weaknesses are not the same. But their variety introduces a physical and technical complexity that requires expertise and processing to extract the information and signals that can be relevant for decision-making. Their measurement technologies and production processes carry errors or uncertainties that can result in apparent inconsistencies that require specific analysis.

We will not develop here full documentation on the different data types but just introduce some characteristics to highlight, the diversity, complementarity, and complexity of the data preparation that is needed to hide the specificities of the measurement process to the users and then better highlight their different assets. Some data, such as the SAR data, are described in more detail because they were critical to the innovation of some e-shape pilots.

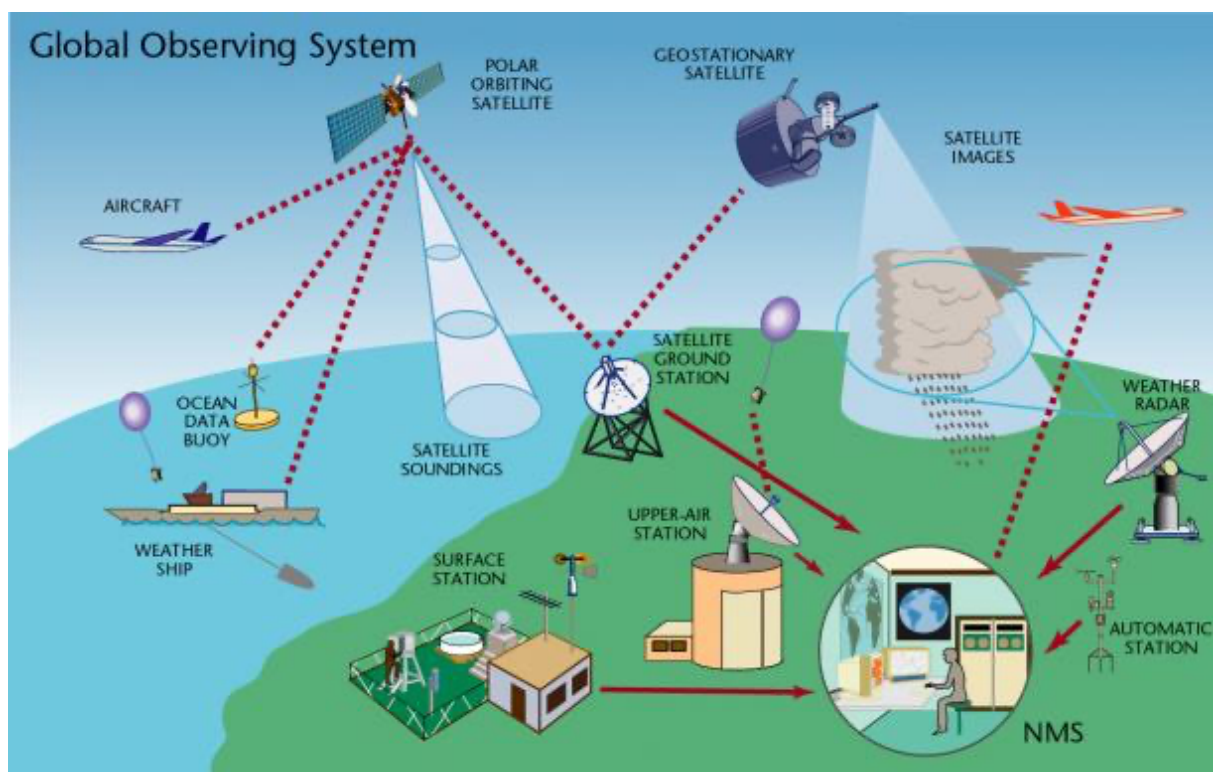


Figure 10 : Global Observing system (source WMO: <https://public.wmo.int/en/programmes/global-observing-system>)

4.2.1.1 Remote Sensing

Remote sensing refers to information about the Earth's surface, atmosphere, and other physical properties using sensors that are not in direct contact with the object or area being studied. Instead, remote sensing uses various instruments and techniques to detect and measure the reflection or emission of electromagnetic radiation from the object or area of interest. Remote sensing developed in the second half of the 20th century and has opened the way to tremendous progress thanks to its global coverage, its regular reliable data delivery, and its quick progress in terms of spatial resolution and temporal frequency.

Remote sensing data can be collected by a variety of sensors, including satellites, airplanes, and ground-based instruments. These sensors can capture data at different wavelengths, such as visible light, infrared, or microwave radiation, allowing scientists to study various features of the Earth's surface, such as land cover, vegetation, soil moisture, and temperature.

Seeing what happens through the clouds is not possible for many sensors but Synthetic Aperture Radar - SAR- can do it, bringing forward a revolutionary potential. Two e-shape pilots have explored the use of SAR data: Assessing Geo-hazard vulnerability of Cities and Critical Infrastructures in the Disaster Showcase and Merging offshore wind products in the Energy Showcase.

SAR images are available since 1992 with the deployment of the ERS-1 (European Remote Sensing-1) satellite enabling new types of remote sensing usages. These data have been continuously evolving, increasing both in spatial and temporal resolution and using different signal wavelengths (L, C, and X bands). ESA and JAXA launched the first C and L band satellites, ERS-1 and JERS-1 (Japanese Earth Resources Satellite), followed by ERS-2 and RADARSAT-1 (ESA and CSA respectively). The second generation included new acquisition modes and improved signal quality. RADARSAT-2 (CSA), ENVISAT (ESA) in the C-band, and ALOS-1 (JAXA) with L-band were launched in the 2001-2007 period. The next period was characterized by the X-band constellations launched by DLR and ASI (Germany and Italy national space agencies). These satellites continue providing very high-resolution images with a high revisit period. The last generation, ESA Sentinel-1, combines high revisiting times, large coverage, and medium-high spatial resolution taking advantage of the TOPS sensor. Sentinel-1 also provides free accessible worldwide data.

Table 3: Main characteristics of the available SAR satellites and constellations

Satellite	Wavelength Band	Incidence angle (°)	Spatial Resolution (m)	Image coverage (km)	Repeating cycle (days)	Space Agency
ERS-1	C	23,5	30	100x100	35	ESA
ERS-2	C	23,5	30	100x100	35	ESA
JERS-1	L	35	18	75x75	44	JAXA
RADARSAT	C	37	28	100x100	24	CSA
ENVISAT	C	21	28	100x100	35	ESA
RADARSAT-2	C	37	28	100x100	23	CSA
TERRASAR-X	X	20-45	3	30x30	11	DLR
COSMO-SkyMed	X	25-50	3	40x40	3	ASI
Sentinel-1	C	20-45	5x20	250x250	6	ESA

4.2.1.2 Geostationary satellites and polar orbital satellites

Satellites are platforms equipped with sensors. They can be geostationary ("Stationary relatively to the Earth") in which case they always observe the same area in time. These Satellites are located at 35,800 km from Earth on a trajectory located at the level of the Equator where the resolution is optimal, and the quality of the data lowers when going towards the poles which cannot really be observed with the geostationary satellites. This is the case for the Meteorological Satellites Meteosat, Goes E and Goes W, GMS, and INSAT. Currently, they provide data collection every 15 to 30 minutes that gather the measurements made scanning from South to North or from North to South over this period of time but each "line" of the scan has a different timestamp: the global collection timestamp is a convention, it can be the time of the start or the end of the scan. The satellites can also be polar orbital and turn around the earth at a lower altitude in a plane a bit inclined relatively to the axe of the poles (693 km for Sentinel 1, 786 km for Sentinel 2, 814 km for Sentinel 3 for instance). They are then heliosynchrone moving relatively to the Earth at the same speed as the Sun. They then observe smaller areas with better resolution data changing with time and they provide data twice a day for the same area. They provide valuable data over the poles where the measurement is very frequent. Sentinels, METOP, NOAA, PROBA-V, ERS... are polar orbital satellites.

The data production process delivers datasets gathering data over a given coverage with a conventional time stamp. This is why building a rapid time series out of Satellite data has been a challenge for a long time that is now addressed with the concept of Data Cubes that enable data exploration in several dimensions with good performances. Sensors delivering comparable measurements can be available on different satellites. Their data can be compiled as a mosaic in space and time after some calibration.

Satellites have a life duration of up to 15 years. One satellite mission is often covered by several satellites providing some redundancy in case of problems that hopefully are not frequent. But it can happen: Sentinel 1B mission has been stopped mid of 2022, during the e-shape project.

The assets of satellite data are that they have global coverage, have a high level of data quality, they are delivered very frequently and in a reliable way. The major challenges are that they deliver huge volumes of data complex to handle as raw data that has to be pre-processed by satellite sensors experts to be usable by other EO stakeholders. This is why the Committee on Earth Observation Satellites - CEOS - community has defined the concept of Analysis Ready Data - ARD - introduced as a new type of valuable data below. The first Earth Observation Satellite image was captured in 1959.

Resources on Satellite missions:

- Earth Monitoring Satellite Mission table from "MOOC to bring AI and Copernicus data together": [https://ugc.futurelearn.com/uploads/files/3b/1e/3b1e8150-8807-483a-aba2-118f1becf806/AI for Earth Monitoring Satellite Mission Table.pdf](https://ugc.futurelearn.com/uploads/files/3b/1e/3b1e8150-8807-483a-aba2-118f1becf806/AI_for_Earth_Monitoring_Satellite_Mission_Table.pdf)

Box 4-1: Satellites used by Pilots. Lessons learnt from the Showcase [MyEcosystem](#) pilot [mySPACE](#) (better monitoring climate drivers in 25 protected areas).

The remote sensing dataset exploited in [mySPACE](#) consists of a time series of Sentinel-2 images to provide the essential variables depending on time such as Hydroperiod, Vegetation Phenology, Gross Primary Production, and Snow Cover Duration.

Both the Sentinel-2 satellites carry onboard an optical payload that detects 13 bands of the electromagnetic spectrum spanning from the visible to the short infrared (SWIR) wavelengths at spatial resolutions of 10 m, 20 m, and 60 m with a revisit time of 2–3 days at mid-latitudes.

The high temporal resolution allows one to focus on the time-series aspect of the data and to analyse dynamic processes with a spatial resolution which has proven to be sufficient for the majority of the selected use cases.

Moreover, the availability of spectral bands spanning the optical range can be fruitfully used to extract spectral indices related to vegetation, water, and snow properties.

Lessons learned from the pilot:

- The essential variables extracted in [mySPACE](#) are useful for end users, such as Protected Areas managers, but an additional step forward that these data offer is the study of the interactions among these variables that help to better understand ecosystem processes and functions

4.2.1.3 In situ data

In contrast to Remote Sensing, in situ data refers to measurements and observations that are taken directly from the natural environment or system being studied. This can include measurements of physical, chemical, or biological parameters such as temperature, salinity, pH, dissolved oxygen, or the abundance of a particular species in a given area. In situ data is often collected using instruments and sensors that are deployed directly in the environment, such as buoys, moorings, or Autonomous Underwater Vehicles (AUVs). This allows for real-time or near-real-time monitoring of environmental conditions and can provide valuable information for scientific research, resource management, and environmental monitoring. It can also be collected by humans, via conventional observations such as Meteorological measurements or via Crowdsourcing Citizen science.

In situ data measures very localized data, in time and space, compared to the areas of interest or the dimensions of the features of interest. It is collected in a location at a certain time as well as in Time series over one point of measurement. Datasets can also gather measurements for a certain time over a collection of points in a predefined area, such as a country for instance.

The World Meteorological Organization - WMO - runs an activity to identify and recognize centennial observing Stations all over the globe (<https://public.wmo.int/en/our-mandate/what-we-do/observations/centennial-observing-stations>) building long series of datasets that can be used for instance to calibrate and validate climate models or climate impacts models.

Long-term observation networks and research infrastructures on a European scale are addressing the aspect of providing continuous and quality-controlled data. This is adopting also the FAIR principles to make in-situ data Findable, Accessible, Interoperable, and Reusable focusing on persistent identifiers, common formats, and consistent and high-quality metadata. Examples of long-term research

infrastructures in the ecosystem and biodiversity domain are [eLTER RI](#), [ICOS](#), [AnaEE](#), or [Danubius RI](#) which all build up and operate data portals with open data licenses.

The very long series is an asset of in situ observations. Their challenges are the fragmentation of the networks, the diversity of sensors in space and time, the missing data, and the fact that they are very irregularly distributed in space (many in developed countries on land, very few in the oceans, remote access areas, or in developing countries).

In-situ data is critical to initialize models, calibrate and validate remote sensing and numerical model products, and run Artificial Intelligence, Model Learning algorithms, or any sophisticated new technologies. The more sophisticated the technology, the bigger the risk is to be unable to identify that the input data were not relevant quantity or quality.

Box 4-2: In situ data. Lessons learnt from the Showcase [MyEcosystem](#).

In situ data. With DEIMS-SDR and EcoSense further developed and refined by pilot [mySITE](#), the mobilisation of datasets by the means of metadata was fostered. The [mySite](#) pilot focused on the mobilisation of in-situ data relevant for EO data pipelines required by [MyEcosystem](#) [mySpace](#) pilot for a number of pilot sites to test and develop the required workflows. All test sites are related to the European LTER network providing a legacy of observation data but still require efforts to harmonise documentation and access to data. [mySITE](#) focused on in-situ observation data e.g. Net Ecosystem Exchange, Land cover and habitats, phenology, water level, leaf area index, canopy height, breast height diameter as well as on biodiversity characteristics. If not available elsewhere, in-situ datasets have been collected and documented using DEIMS-SDR as well as B2SHARE as a repository.

Lessons learned

- data provision and sharing are still hampered by different policies applied and data management practices used. Further efforts to mobilise and harmonise data from long-term observation are needed
- means of automatically generating interoperable metadata records need to be enhanced

References:

- DEIMS-SDR (Dynamic Ecological Information Management System - Site and dataset registry) <https://deims.org/>
- B2SHARE. <https://b2share.eudat.eu/>

4.2.1.4 Citizen science data

One specific kind of in-situ data is Citizen science or crowdsourced data, which are measured and provided voluntarily by citizens. Long seen by scientists as dubious, it is now developing as a new source of impacting data and as a new science itself as it requires the development of new data management and processing techniques. The benefit is that it allows capturing data that cannot be easily captured by sensors (ex: the number of butterflies of a specific type in an area), that it is inexpensive and scalable, that citizens are happy to engage on topics that are of interest for them or just to help science. There is a risk of volunteer or non-volunteer bad quality inputs, but given a large enough sample size, this can be mitigated with some statistics or automatic controls. In addition, citizen-science observations also have the risk to include a bias towards very prominent species or features in the landscape (e.g. paddy rice fields), which may impact the applications for which these data are used.

Within e-shape, there were some extreme events impacting the collection of data through citizen science, namely conflicts within countries, and the COVID pandemic. While the Covid pandemic halted many professional activities, including field trials, it also brought a large surge in outdoor activities for people that were confined to their residence. This surge was one of the motivations for the CropObserve app development, as an unprecedented number people were walking frequently through the agricultural production areas, making them the perfect data collectors to capture both the crop types as well as the management activities on the field. Escalating conflicts, on the other hand, proved to be a real challenge, also for crowdsourced data collection. Both the civil war in Ethiopia and the Ukrainian conflict had a severe impact on field data availability, and the roll-out of data collection tools in the field.

Citizen science data also raised the difficult issue of data privacy (See paragraph on Data Privacy)

Box 4-3: Citizen Data Science. Lessons learnt from the [GEOGLAM](#) Pilot - The [Food Security and Sustainable Agriculture](#) showcase example

The [GEOGLAM](#) Pilot has developed the [CropObserve](#) app, to circumvent the issue of limited agricultural reference data in EO service development. CropObserve is a mobile application that was developed to allow anyone to observe agricultural fields and crops anywhere on Earth. The app is focused on collecting crop type (e.g., cereals), phenological stage (e.g., flowering), visible damage (e.g., frost), and management practices (e.g., harvest), with different levels of complexity to ensure that even non-experts can provide useful and reliable information on agricultural production areas. It is strongly encouraged to include geo-tagged photos, to facilitate a more rigorous verification of the observations. After data collection, the data undergo quality checks before being ingested in the [AgroStac](#) data portal, to ensure that the data become FAIR compliant. Since its release, the uptake has been tremendous, with around 1,300 observations in 2021, and 1,900 observations in 2022. Although the majority of these observations are from Belgium (3,388), where specific uptake initiatives were organized with universities and local governmental agencies, the list of countries is expanding with observations in Austria (30), the Netherlands (51), and Argentina (105). The observations collected with CropObserve enabled the development of more rigorous and scalable EO methodologies on crop calendars. As a result, the crop calendar methodologies became applicable at scale, and are directly used in the development of the Copernicus Cropland High-Resolution layers.

4.2.1.5 Numerical models

A Numerical Model is a tool that facilitates the evaluation of a system's status, its behaviour, and its effects on its related systems or environment. Numerical modelling is a technique used by EO practitioners that incorporates the most up-to-date theories that attempt to formulate Earth-related status, processes impacts, or behaviour. Numerical models are used in most of the industrial and research domains. Prediction models, such as weather or hydrological models, attempting to predict the behaviour of a system in the future (the so-called scenario), allow one to anticipate and get prepared for the impacts. They can be instrumental to save lives, protecting goods, for societal smart development, or for economic advantage. They can be global and simulate any area of interest including the globe, or local to focus with a finer mesh over a restricted area. They can run over past periods of time or forecast future behaviours of the feature of interest. Several models can interact to downscale the results from a global scale to a local scale or to couple different domains that interact physically in the real world such as the atmosphere and oceans.

Numerical models are complex to develop, they can need a lot of computing power and resources requiring the sometimes most powerful computers. Running these computers is very expensive in terms of electricity and staff. Many numerical models use Earth Observation data to initialize their physical

equations. The first steps of a model, called data assimilation, consist in ingesting these observations into the physical representation of the model making a numerical image of the status of the system named "analysis" in meteorology for instance. This can be considered as a simulated observation. Numerical models can integrate knowledge from different domains such as, for example, atmospheric physics, chemistry, biology, and hydrology.

Earth observations are instrumental to initialize, calibrate, validate, and challenge the numerical models outputs. Numerical models can fill the gaps in 3D and time for other earth observations. The outputs of numerical models can be integrated with statistical emulators that approximate their complexity when their computational load limits any sensitivity and uncertainty analysis. Statistical emulators can be used to develop user-friendly tools for scenario analysis.

Box 4-4: Numerical models Lessons learnt from the [EO-based surveillance of Mercury pollution](#) (Minamata Convention) Pilot - [Health Surveillance](#) showcase

The Pilot [EO-based surveillance of Mercury pollution](#) (Minamata Convention) has developed a modular tool HERMES (<https://gkhub.earthobservations.org/packages/2wxxd-w9009>) that is designed to connect a statistical emulator of a Chemical Transport Model (CTM) with a Bio-geochemical Model (BGCM) and in downstream a trophic model. The CTM can simulate a high-resolution fate of mercury (Hg) in the atmosphere from the emission source to the final receptors, and therefore allow for a source-receptor assessment. Simulations of the Hg atmospheric cycle by CTMs have a temporal limit (a few years), due to the fact that Hg exchange at the interface of the atmosphere with other compartments (soil and oceans) is poorly characterized. To override the limitation a BGCM is coupled with CTM to simulate the Hg exchange between Earth System compartments on a time scale of decades and centuries. The basic idea behind the development of the statistical emulator is to run a number of anthropogenic emission control scenarios and to analyse the response of the system in order to find a predictable pattern. If such a pattern exists, it can be used to simulate the response of the system for all other possible emission scenarios, thus saving time and computational costs. The fitting model used to approximate the response function might have a reasonable error, which magnitudes clearly depend on the goodness of the model itself.

Box 4-5: Numerical models. Lessons learnt from [EYWA - Early Warning System for Mosquito-Borne Diseases](#) Pilot - [Health Surveillance](#) showcase

The Pilot EYWA ([Early Warning System for Mosquito-borne diseases](#)) has developed the MAMOTH mosquito abundance prediction tool, a data-driven machine learning model (<https://gkhub.earthobservations.org/packages/etc3g-a6t86>). This model is being trained on time-series of in-situ historical entomological data collections, which are augmented by Earth Observation data representing the conditions on the ground at the time of the collection. Specifically remote sensing data are used to create vegetation, moisture, water, and build-up proxies of the environment as well as statistical features from the meteorological conditions (rainfall, land surface temperature) and the geomorphological landscape of the trapping sites. This large dataset is used to generate a model that given the specific conditions at any place and time, generates a prediction for the expected mosquito populations at the upcoming time (15-30 days in the future). Such predictive information is used by vector control companies and health authorities to plan larviciding actions and guide awareness campaigns in at-risk regions.

Box 4-6: Numerical models. Lessons learnt from the [Health Surveillance Air Quality Pilot - Health Surveillance](#) showcase

The Pilot HSAQ ([Health Surveillance Air Quality](#)) has incorporated several developments on health-related air quality products in the urban environment (<https://gkhub.earthobservations.org/packages/jaf2d-7bj57>). In particular, the Athens component is designed to connect a Chemistry Transport Model (CTM) with a meteorological model and Earth Observation (mainly Copernicus) data representing land use/cover, anthropogenic emissions, boundary air pollution concentrations and population density. The CTM can simulate a high-resolution fate of atmospheric pollutants (NO_x, SO₂, CO, O₃, PM_{2.5}, PM₁₀, NMVOCs) in the atmosphere from emission sources to population exposure and therefore allow for health impact assessments. Simulations of atmospheric pollutants have been performed for an indicative, contemporary, restriction-free year and are presented on a monthly basis. The basic idea behind this development and service is to exploit intra-urban (100m spatial resolution) CTM results towards health-related products, i.e. the air quality index, the number of exceedance days of the latest WHO limits, and the amount of population exposed to air concentrations above safe limits. Such added value atmospheric information can be used by health authorities to support the establishment of a National Environmental Health Observatory.

4.2.1.6 Ground Radar data

In meteorology, precipitation radar data are very used to for precipitation short term forecasts and to issue weather warnings. The advantage of weather radar is that it provides 3-D observations at a high spatial and temporal resolution and with large coverage large compared to in situ, and small compared to satellites). But these data require specific expertise to be used efficiently as the beam of the radar is tangent to the earth and measures reflectivities, velocities... at different altitudes depending on the distance to the radar, the images can sometimes be affected by various artifacts and errors, such as attenuation, clutter, and noise due to some specific propagation of the beam...Radar data have to be interpreted in conjunction with other data and by experts.

They remain critical data for weather and hydrological forecasts.

In e-shape only the Pilot 2: [GEOSS for Disasters in Urban Environment](#) (improved resilience of cities, infrastructure and ecosystems to disasters) of Showcase 6: [Disasters Resilience](#) have used these data.

4.2.1.7 Lidar data

Lidar (Light Detection and Ranging) data is obtained using a remote sensing technique that measures the distance between a sensor and a target surface by sending laser pulses and measuring the time it takes for the light to return to the sensor. Lidar data can have several characteristics or parameters that describe the properties of the target surface and the surrounding environment. Lidar data is widely used in various applications, including topographic mapping, urban planning, forestry, flood modelling, infrastructure assessment, and many others.

In e-shape the Pilot 1: [Data for Detection, Discrimination and Distribution \(4D\) of Volcanic ash](#) of Showcase 6: [Disasters Resilience](#) have used these data.

4.2.1.8 Internet of Things - IoT- data

Internet of Things (IoT) data refers to data generated by various interconnected devices, sensors, and systems that are part of the IoT ecosystem. IoT data can have several characteristics, which include

Volume, Velocity, Variety, Veracity, Variety of sources, Contextual information such as location, time, and other metadata, Concerns about Security and privacy, Scalability, Data integration, Time sensitivity.

More and more IoT data are produced, applying to different domains and they are at the origin of the new emphasis on Data Spaces. Data Spaces concept is still very recent. It aims at federating several sources of IoT with a strong focus on Sovereignty, Trust and Quality. At the moment it is very pushed by the private market. These data will probably take more and more importance in the future but still have a lot of challenges to address.

Showcase 1: [Food Security and Sustainable Agriculture](#) Pilot 5: [Linking EO and Farm IoT for Automated Decision Support](#) and Pilot2: [EU-CAP Support](#) (improved efficacy of implementing CAP and its underlying principles of environmental stewardship) have used IoT Data.

4.2.2 Value-added Data products, concepts or services:

4.2.2.1 Historical data and Time series

Data analysis is mainly done in space and then over time or over time in a specific location. Historical data are the data collected and stored through time. They are critical to understanding the behaviour, variability, and evolution of resources so that environmental issues and impacts on societal challenges can be planned accordingly. Long-time series of historical data enable the analysis of the normal (eco)-system condition and behaviour, and consequently deviations and variability from this behaviour, including extreme values, trends towards new means, new extremes, new impacts, and risks.

Time series and the need for efficient analysis in whatever dimension are one of the reasons for the Data Cubes concept and technologies.

4.2.2.2 Reanalysis, Forecast, Seasonal and Sub-seasonal climate data

Reanalysis data are a "blend of observations with past short-range weather forecasts rerun with modern weather forecasting models. They are globally complete and consistent in time and are sometimes referred to as 'maps without gaps'. They provide the most complete picture currently possible of past weather and climate". In other terms, it is the results of Earth Observations data assimilation by the model (in situ, remote sensing..) over the model grid made with a recent model providing a solution to the gaps, the irregularities, and the errors of the Earth Observations. The reanalysis data is produced with intense quality control activities with this recent model over many years of past data resulting in a complete and homogeneous data cube of data from the Earth surface to the top of the atmosphere for the most common parameters. These data are freely available through the [C3S Climate Data Store](#) Copernicus Service. "The most recent ECMWF reanalysis dataset is the ERA5 Back Extension, a preliminary release to be further updated in 2021 that provides data from 1950-1978 and sits alongside the main established ERA 5 dataset (1979 to present day)". Other major climate data producers such as NOAA also produce reanalysis data.

A forecast is about making predictions about the future. It is very popular for weather and climate but forecasts are also used Energy production and consumption, for all stages of agriculture, for health, for biodiversity, water resources, demography, finance...Forecasts are made with many different types of models and support all types of decisions. The forecasts are never perfect, because if they were perfect, they would be never wrong and would have no limits in time. Unfortunately, we know that they can fail and that the uncertainty of the forecasts grows rapidly with the range of the forecasts. This is why the expertise of the forecasters, based on their knowledge of the model and their experience, is to regularly assess the Earth Observations (reality) with the model forecasts (simulations) to improve if possible the timing of the impacts.

Seasonal forecasts are not as accurate as forecasts and rather give trends over a larger region and for a longer period of time, typically a season.

Sub-seasonal forecasts, as defined by FMI in its Climate pilots, cover a period of six weeks, between forecasts and seasonal forecasts, which is a period of time for which a lot of businesses have to make decisions impacting their activities. This is the case for the [Seasonal Preparedness](#) pilot presented below.

Box 4-7: Sub-seasonal forecasts for seasonal preparedness in extreme climates. Lessons learnt by FMI.

The forecast data used by the Pilot [Seasonal Preparedness](#) (improved transportation safety in extreme climates and tourism impact indicators) of the showcase [Climate](#) in the development of tailored sub-seasonal and seasonal climate products for tire companies (by FMI) and operational production are the Extended-Range Forecasts (ERFs) and Long-Range Forecasts (LRFs) from the European Centre for Medium-Range Weather Forecasts (ECMWF).

The extended-range (or sub-seasonal) forecasts are provided by the Ensemble Prediction System (EPS) of the European Centre for Medium-Range Weather Forecasts (ECMWF, 2016) and are used in the production of the sub-seasonal (6-week) outlooks. EPS includes 51 ensemble members, and the weather forecasts are extended up to 46 days twice a week, on Mondays and Thursdays. The horizontal resolution of the forecast is 0.4° (~36 km) and the surface-based data is available 6-hourly. Besides real-time forecasts, retrospective forecasts (re-forecasts) are also created for the past 20 years and used for quality assessment and calibration of the model variables used. Since the ERF data is not available in the C3S Climate Data Store, it can be accessed from ECMWF through the Meteorological Archival and Retrieval System (MARS) for the development phase and through the ECMWF dissemination (used in the operational service). The variables used in the outlooks' development and production are the 2m temperature, snow depth, and snow density.

The long-range (or seasonal) forecasts used in the production of seasonal climate outlooks are provided by the SEAS5 seasonal forecast system of ECMWF (Johnson et al. 2019) accessed from the C3S Climate Data Store, where it is updated on the 13th of every month. The spatial resolution of the data is 1° x 1°, the model system includes 51 ensemble members for real-time forecasts. Re-forecasts are also updated together with real-time forecasts in the same resolution for 25 ensemble members. Re-forecast data for the period 1993-2016 was used in the post-processing of the variables. The variables used are the 2m temperature, snow depth, and snow density.

The sub-seasonal and seasonal forecast variables used in the development of the service are quality assessed and the systematical biases from raw data are removed through bias adjustment methods.

ERA5 reanalysis (Hersbach, 2020) data were used as reference data in the calibration of ERF and LRF variables and forecast verification. It is accessible from C3S Climate Data Store and it is a component of the production flow within the operation service for both seasonal and sub-seasonal climate outlooks.

The service implemented for the tourism sector (by AA) uses the same SEAS5 seasonal forecast model and ERA5 reanalysis. The selected parameters are 2m temperature, relative humidity, wind speed, and precipitation.

References on Reanalysis:

- ECMWF Fact Sheet: Reanalysis: <https://www.ecmwf.int/en/about/media-centre/focus/2020/fact-sheet-reanalysis>

- NOAA Website: <https://www.ncei.noaa.gov/products/weather-climate-models/reanalysis>

References on seasonal forecasts:

- FMI Seasonal forecasts: <https://seasonal.fmi.fi/e-shape/helsinki/>

4.2.2.3 Analysis Ready Data

CEOS has been the first to formalize the concept of Analysis Ready Data (ARD) as "*CEOS Analysis Ready Data (CEOS-ARD) are satellite data that have been processed to a minimum set of requirements and organized into a form that allows immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets.*"

The goal is to make satellite data more usable by users who are not high experts in raw data satellite technologies and cannot download big amounts of data due to bandwidth or infrastructure limitations. It also avoids running many times the same basic processing that is needed by all. ARD is a major enabler in mainstreaming the use of satellite data.

The OGC Engineering report "[OGC Testbed-16: Analysis Ready Data Engineering Report](#)" (OGC 20-041) generalizes the ARD concept and studies its implications for the OGC Standards baseline. In particular, the Report analyses how modern federated data processing architectures applying data cubes and Docker packages can take advantage of the existence of ARD. The main characteristics of the CEOS ARD for Land (CARD4L) can be extended to other dataset sources including in-situ measurements. More work will follow on this topic in the OGC community.

With CEOS support, OGC and ISO/TC 211 have a joint new project to consider this specification and whether it can be extended to describe analysis-ready data from sources other than satellites. The ISO new work item proposal will be voted on soon as ISO 19176-1 Geographic information "Analysis Ready Data — Part 1: Framework and Fundamentals"; OGC will start a vote in 2023 on establishing an Analysis Ready Data Standards Working Group. The plan is to develop the ARD standards by using CEOS ARD specs as the base, incorporating relevant content from existing OGC and ISO standards.

Box 4-8: [The need for consistent time series - an example from the Vegetation-Index Crop-Insurance in Ethiopia pilot in the Food Security and Sustainable Agriculture showcase.](#)

Long time series often consist of data coming from multiple satellites, that were launched in succession. This ensures that the Earth's processes can be monitored in a continuous manner, which in turn provides valuable information on how these processes are changing through time and space. However, these different satellites inevitably come with different technical specifications, and discrepancies exist between the datasets. For this reason, consecutive missions ideally include an overlapping period where data from both satellites is available, to enable the quantification and characterization of these differences.

For continuous data monitoring and data offering, it is essential that this data alignment is done prior to the decommissioning of the former satellite. If not, operational monitoring services cannot guarantee data continuity, which in turn severely jeopardizes the service offering and the contractual obligations that go with it.

A concrete example of this was shown in the index insurance scheme in Ethiopia, where an insurance product was sold to smallholder farmers in Ethiopia. It was built and calibrated on data from the PROBA-V satellite, which was launched in May 2013 and reached the end of its operational lifetime end of June 2020. However, in late November 2020, it was still using the non-corrected PROBA-V products for Africa (the processing line was temporarily re-established for this purpose) because

there was no alternative : the Sentinel-3 SYNERGY data from Copernicus that was planned to be the successor of the Proba-V data, and which was foreseen to be included as the data source for the insurance product, had been rejected by the pilot since the atmospheric correction was of poor quality, and a misalignment existed between the NDVI data of the Proba-V archive and the Sentinel-3 data; this proved to be specifically true for the highlands of Ethiopia. This misalignment had huge repercussions on the sales of the insurance product for the 2021 season, and alternative data sources were needed.

On a specific request to ensure operational continuity, past BRDF-adjusted 1 km resolution data based on Spot-VGT and Proba-V were made available (courtesy of VITO), which enabled the pilot to redefine the insurance product based on this new archive covering 20 years (1999 onwards), which provides the variability (impacts by weather) as needed to properly establish the thresholds by zone of the lower extreme NDVI-ranges. The pilot (UTwente and Mekelle University) processed the past BRDF-data in anticipation of, and preparation for continuity of the VICI insurance in 2021. During the insured part of the growing season, BRDF-adjusted data at 300 meters resolution, derived from Sentinel-3 imagery, was used, thus enabling the continued offering of the insurance products.

Box 4-9: “ARD online” as an enabler of javascript-based EO Apps

FMI has produced ARD data according to CEOS specifications from Sentinel-1 and -2 into dekad mosaics over all of Finland available 3 times a month. Very practically seen, ARD combined with open distributed cloud technologies like SpatioTemporal Asset Catalogues and Cloud-Optimized GeoTiffs (COG) can allow geospatial web-services to be run from web-browsers. Which means that developing EO apps is only a javascript writing process for openly available ARD online. This has been demonstrated with an app for analysing forest damage: <https://tuulituhohaukka.out.ocp.fmi.fi/>. Some of these Cloud Optimized GeoTiff -COG -layers are also being used in the Forestry conditions (more efficient forestry operations with lower environmental impact and carbon emissions) Pilot service from Climate Showcase.

Lessons learned on ARD:

- The concept of ARD is fully adopted by the users and several pilots have expressed the need to have more ARD products available to save time, and processing costs and benefit from the upstream expertise.
- ARD products are defined by data providers and should converge, at some point, with Essential Variables being defined by the users 'communities.

References on ARD:

- CEOS Analysis Ready Data : <https://ceos.org/ard/>
- OGC Testbed-16: Analysis Ready Data Engineering Report: <https://docs.ogc.org/per/20-041.html>
- News from ISO/TC 211 Geographic information/Geomatics Published 2023-03-03 <https://committee.iso.org/home/tc21>
- OGC Analysis Ready Data Standards Working Group Charter under public review, 29 November 2022, <https://www.ogc.org/news/public-comment-requested-on-draft-charter-for-new-ogc-analysis-ready-data-standards-working-group/>

4.2.3 The Copernicus Data and Services

Copernicus is the European Union's Earth observation programme, looking at our planet and its environment to benefit all European citizens. Copernicus is comprised of three components: In Situ, Space, and Services.

- Copernicus In situ: <https://insitu.copernicus.eu/about>
- Copernicus Space: <https://www.copernicus.eu/en/copernicus-satellite-data-access>

The 6 information services draw from satellite Earth Observation and in-situ (non-space) data.(source <https://www.copernicus.eu/en/about-copernicus> and <https://insitu.copernicus.eu/about>)

- Copernicus Information Services (Figure 11):
 - Copernicus Atmosphere Monitoring Service - CAMS:
<https://www.copernicus.eu/services/atmosphere>
 - Copernicus Land Monitoring Service - CLMS:
<https://www.copernicus.eu/services/land>
 - the Emergency Management Service – Mapping - EMS⁴:
<https://www.copernicus.eu/services/emergency>;
 - Copernicus Marine Environment Monitoring Service- CMEMS:
<https://www.copernicus.eu/services/marine>
 - Security:
<https://www.copernicus.eu/services/security>
 - Climate Change Service - C3S:
<https://www.copernicus.eu/services/climate-change>



Figure 11: Copernicus Services, a user driven approach

⁴ The Emergency and Management Service can be activated only by designated authorized users.

All these data and services are free of charge and their license only requests basic crediting. They are fully operational 24/7.

The Copernicus Programme implements a user-centric development process and organizes many events to strengthen the connection with its users' community and share knowledge on their products and services.

Lessons learned on Copernicus data:

The Copernicus data and services are very valuable and 34 of the 37 e-shape pilots used Copernicus data (See: Annex 4: Copernicus Services used by the e-shape pilots).

4.2.4 Essential Variables

The laws of Physics, the vegetation development process, the epidemics development models, the Sustainable Development Goals, and other rules guiding nature's or society's evolution guide scientists on the choice for the physical quantities they need to measure or assess. Different communities collaborate around the world to define the Essential Variables critical to monitor their features of interest and support the decisions in their domain.

Some domains have already agreed on a baseline of variables: The Global Climate Observing System (GCOS) has identified a set of 54 atmosphere, land, and ocean variables as Essential Climate Variables (ECVs); United Nations Educational, Scientific and Cultural Organization's (UNESCO) Global Ocean Observing System (GOOS) has defined Essential Ocean Variables (EOVs); The Group on Earth Observations Biodiversity Observation Network (GEO BON) has identified six classes of Essential Biodiversity Variables (EBVs)—one for genetics, two for species, and three for communities and ecosystems, and several EBVs have initially been identified within each class. They are complemented by Essential Geodiversity Variables (EGV) important in the availability of natural resources and connections with other essential variables. GEO's Global Agricultural Monitoring (GEOGLAM) has defined Essential Agriculture Variables (EAVs), translated them into EO Data Requirements, and is monitoring their developments via the definition of priorities and updated status (see <https://agvariables.org/>).

As a matter of fact, identifying the Essential Variables is a major step in the translation from Needs to Data Requirements. Data Requirements then allow connecting the Data Demand to the Data offer. We can then really talk about an EO Market, driven, as any other Market, by the offer and the demand.

Considering the Data Offer, it relies on a portfolio of various types of data with different strengths and weaknesses that have to be considered as complementary. It also relies on pre-processed products that hide the complexity of the measurement process to build value-added data products that can be reusable by many users. ARD products can be based on a single type of EO data (e.g. Remote sensing) or on the fusion of several (ex: ECMWF Reanalysis - ERA 5).

A new approach for Essential Variables is also forming to understand better the downstream uptake of data in the Shared Arctic Variables.

References on Essential Variables:

- GEO community activity report "Mainstreaming Essential Variables across GEO": https://www.researchgate.net/publication/367361242_GEO_community_activity_report_Mainstreaming_EVs_across_GEO/link/63cf8e95e922c50e99bb566c/download
- [GEOGLAM](https://agvariables.org/) Essential Agriculture Variables & Agricultural Indicators for GEOGLAM: <https://agvariables.org/>
- The SMURBS community activity report "Essential Variables and Indicators", a collaborative effort of the partnership of SMURBS/ERA-PLANET project proposes Essential Urban Variables (EUVs) on the fields of air quality, disasters, and urban growth. The deliverable can be found here: https://drive.google.com/file/d/1v_LrBuc8nP9GR-Tii6SsJ3v3IHZR_b1r/view

Box 4-10: Essential Variables. Testimonial from Showcase [MyEcosystemPilot](#) [myVARIABLE](#) (MLU, UT, SYKE, WR)

The biodiversity crisis we are experiencing requires the establishment of an observation and monitoring system to help us understand where we have the greatest problems, to inform actions to halt and reverse the loss in these places, and to anticipate the impact of our future actions. We need a system analogous to the Global Climate Observing System (GCOS). Underpinning this system are the tremendously successful standards that helped to harmonize the information produced by GCOS. However, establishing standards is one thing, and successfully implementing them is quite another. Achieving universalization of a standard requires hard work to overcome many aspects that are often more human than technical, including inertia, resistance, competition, ignorance, lack of resources, and legal barriers, among other problems. Even if everyone agrees and wants the same standard, it is often difficult to avoid local variation, especially when standards require humans to act uniformly across large expanses of time and space. In trying to develop a standard for the EBVs for biodiversity, we face an additional and particularly complex problem as the nature of the object of our efforts is its intrinsic variability and constant evolution. Instead of trying to reinvent the wheel we stand on the shoulders of the giants of the climate community and adopt and slightly modify the data standard they use for our EBVs!

Lessons learned:

- In science and development, we often strive to be the first and only ones to discover solutions. However, many times the solutions we need are found just around the corner, in other domains and areas. It's important not to close our eyes to these other areas, as they can provide valuable insights and guidance in unexpected development directions.
- One key issue with standards is adoption. If we can leverage the momentum of adoption for a standard that other communities have already achieved, it can make the process faster and more efficient.

4.3 Data quality

4.3.1 Introduction

Reusing data in downstream processing or for any business-critical decision requires trust. Trust and quality are closely linked concepts. When you trust a product or service, you expect it to meet a certain level of quality. On the other hand, when you encounter high-quality products or services, you are more

likely to trust them. The advantage of the term quality over trust is that it is an IT-specific concept that can be defined as the degree to which the data compare to other similar data but also how it fulfills the requirements. This chapter will only review as an example, some of the work related to data quality developed by the e-shape pilots.

4.3.2 Standards for Data and services quality evaluation and assessment and management

ISO 19157:2031 standard provides a framework for evaluating and managing the quality of geospatial data, including spatial data and metadata.

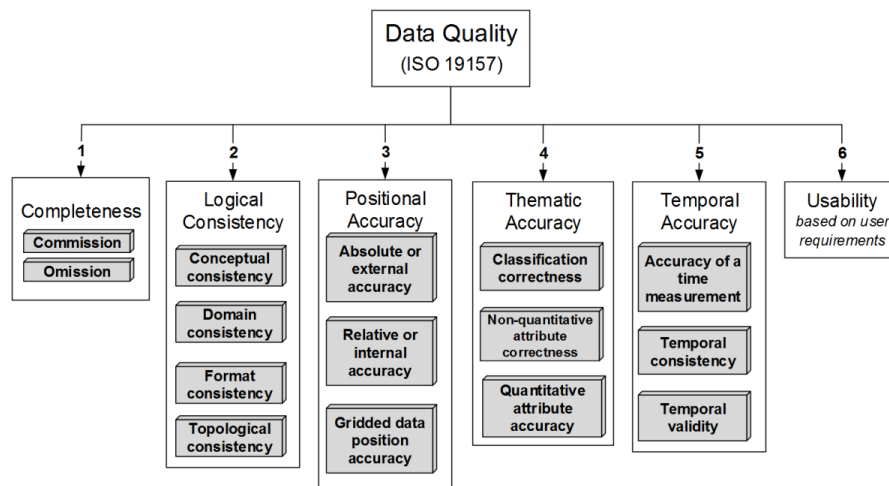


Figure 12:INSPIRE Technical Guidelines use ISO 19157 Geographic Information Data quality. Source: EEA.

ISO 19131 and ISO 19158 can also be of interest in this field. ISO 19131:2007 provides a framework for managing the quality of geospatial processes, including data acquisition, data processing, and data analysis. ISO 19158:2012 is an International Standard developed by the International Organization for Standardization (ISO) that provides guidelines for the quality assurance of data for spatial data infrastructures (SDIs).

4.3.3 Forecast skill assessment

"In the fields of forecasting and prediction, forecast skill or prediction skill is any measure of the accuracy and/or degree of association of prediction to an observation or estimate of the actual value of what is being predicted (formally, the predictand); it may be quantified as a skill score." (source Wikipedia)

This skill can be improved by applying some bias adjustment thanks to new related datasets, statistical models, or methodologies, recalibrating the model outputs. The accuracy can be improved without degrading the global skill. The completeness of the metadata is critical for the reusability of the datasets and needs to provide metadata on the data provenance that will provide all information on the processing applied to data.

Box 4-11: Data quality. Lessons learnt by FMI, ZAMG/Geosphere, DWD across multiple Pilots

FMI: Use of Statistical methods to adjust SEA5 Seasonal forecasts over Finland

The seasonal snow accumulation outlook developed by FMI uses as input the data provided by the SEAS5 seasonal forecast system of ECMWF. The skill of snow data was assessed previously for the whole Finland using the high-resolution re-forecasts provided by the same forecast system for the period 1993-2016 accessed from MARS archive server from the European Centre for Medium-Range Forecast - ECMWF. The variables used were snow depth and snow density. ERA5 re-analysis data accessed from C3S Climate Data Store for the same period as re-forecasts was used in the skill assessment and bias adjustment of snow data. The skill of snow data was assessed for raw forecast ensemble and for the data corrected with empirical quantal mapping, variance, and ensemble model output statistics (EMOS). The evaluation was performed for three lead months using reliability diagrams of aggregated grid points of Finnish land areas and maps of verification measures, i.e. continuous ranked probability skill score (CRPSS) and mean error. Reliability diagrams of snow depth forecasts indicated that all the terciles for all three lead months are usable, varying between perfect and marginally useful for all the initializations except October, when the forecasts are mostly unusable. High CRPSS values that are statistically significant for large areas for lead months 0 and 1 indicate skillful forecasts for Dec-March period. According to skill assessment analysis the EMOS method improved the skill of raw forecasts the most.

The skill of sub-seasonal forecast data used in the development of sub-seasonal forecast outlooks was also evaluated. The temperature, snow depth, snow density, and total precipitation data were used from the Ensemble Prediction System (EPS) of ECMWF in the evaluation. Observations from Helsinki Kaisaniemi for the period 2000-2019 are used in the verification and calibration of variables. Verification was run so far for 2m temperature, post-processing of snow and precipitation data is ongoing.

ZAMG/GeoSphere Austria: Input data for urban climate model simulations

The MUKLIMO_3 model uses various spatial data to describe land surface properties and urban structures. In the pilot were used combined datasets from Copernicus Land Monitoring Services, such as Urban Atlas for land use classification and High-Resolution Layers for tree cover density, and national data on land cover such as LISA and additional data on degree of soil sealing and building height.

Urban climate model in combination with regional climate model scenarios is used to derive climate indices related to heat conditions such as the annual average number of summer days (days with maximum air temperature $\geq 25^{\circ}\text{C}$). The regional climate model simulations from the EURO-CORDEX project provided information on meteorological variables until 2100 under various climate change scenarios, e.g. RCP4.5 and RCP8.5. In the pilot, eight different combinations of bias corrected regional/global climate model output were used to calculate climate indices for each year between 2011 and 2100.

DWD: Skill assessment and observed input data

The skill of climate prediction is evaluated against the skill of the reference prediction climate mean (climatology) observed over the evaluation period. The evaluation period for the seasonal climate predictions is 1991-2020. The skill score of the ensemble mean prediction is determined using the skill score of the mean-squared error between the ensemble mean of the hindcasts and the actual observation (MSESS). The prediction skill score of the probabilistic predictions is determined using the ranked probability skill score (RPSS), which examines to what extent the predicted probabilities of occurrence of the categories (e.g. 'below normal', 'normal', 'above normal') agree with the category actually observed. Furthermore, a bootstrapping method is applied to verify whether the

skill comparison between hindcasts and a reference prediction is subject to accidental variations due to small sample sizes (significance test).

The reference prediction for the skill score is the observed climatology. The observed precipitation climatology corresponds to the 5 km gridded observation data HYRAS (Razafimaharo et al. 2020, Rauthe et al. 2013) and the observed temperature climatology corresponds to a gridded DWD monthly averaged 2 m air temperature dataset over Germany with a spatial resolution of 1 km.

4.3.4 Assessing input data skills versus output data skills

When processing data to improve its quality in any of the above criteria, it is important to assess the resulting data with the same process as the one applied to the input data to control that the processing did not degrade or introduced bias in the data skills.

Box 4-12: Data quality. Lessons learnt by DWD and ZAMG/Geosphere, in the Climate Showcase.

The Urban resilience to extreme weather (better forecast of heat stress, rainfall, and storms in urban areas) Pilot from the Climate Showcase example

DWD increases the spatial resolution of the global climate predictions using the empirical statistical downscaling method EPISODES (Kreienkamp et al. 2019). EPISODES conserves the skill of the downscaled hindcasts for the two variables near-surface (2 m) temperature and precipitation. Thus, the seasonal climate information is available at a higher spatial resolution without losing skill. Furthermore, the output of the statistical downscaling is nearly bias-free, which is, besides the higher spatial resolution, an added value for the climate service (Ostermöller et al. 2021).

Kreienkamp, F., et al., 2019: Evaluation of the empirical-statistical downscaling method EPISODES, Climate Dynamics, 52, 991-1026. <https://doi.org/10.1007/s00382-018-4276-2>

Ostermöller, J., et al., 2021: Downscaling and Evaluation of Seasonal Climate Data for the European Power Sector, Atmosphere, 12(3).

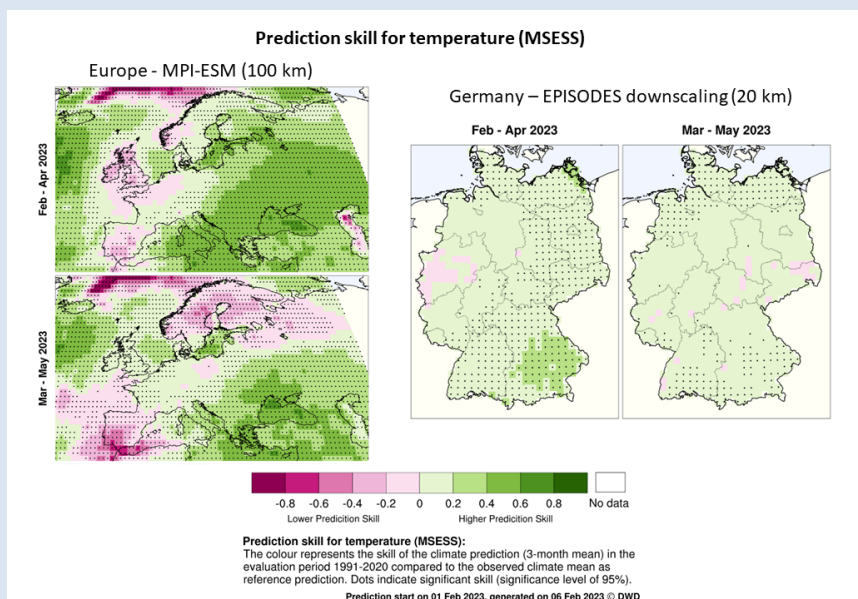


Figure 13 : Prediction skills for temperature (MSESS)

ZAMG/GeoSphere Austria: Downscaling of climate scenarios using urban climate model simulations

The urban climate model simulations " (Sievers et al. 2016) simulate radiation, soil and atmospheric temperature, relative humidity, and wind flow in urban areas on a 3D grid (100 m spatial resolution, vertical resolution 10-100 m). The results for daily minimum and maximum temperature of the model provided the basis for evaluation of the climate indices within the urban area. In combination with regional climate model scenarios (representative for the background climate by applying the cuboid method (Früh et al. 2011; Žuvela-Aloise et al. 2014; Geletič et al. 2019), the model was used to derive climate indices related to heat conditions for future climate. The calculated indices on 100x100 were compared to SPARTACUS/OKTAV climate scenarios for Austria (Chimani et al. 2016) showing more spatial detail, but similar climate signals (see below).

Chimani, B., Heinrich, G., Hofstätter, M., Kerschbaumer, M., Kienberger, S., Leuprecht, A., ... & Salzmann, M. (2016). Endbericht ÖKS15–Klimaszenarien für Österreich-Daten-Methoden-Klimaanalyse. Projektbericht. CCCA Data Centre. <https://data.ccca.ac.at/dataset/a4ec86ca-eeae-4457-b0c7-78eed6b71c05>.

Früh, B., Becker, P., Deutschländer, T., Hessel, J.-D., Kossmann, M., Mieskes, I., Namyslo, J., Roos, M., Sievers, U., Steigerwald, T., Turau, H. und Wienert, U. 2011: Estimation of climate-change impacts on the urban heat load using an urban climate model and regional climate projections. J. Appl. Meteorol. Climatol. 50 (1), 167–184.

Geletic, J., Lehnert, M., Dobrovolny, P., & Zuvela-Aloise, M. 2019: Spatial modeling of summer climate indices based on local climate zones: expected changes in the future climate of Brno, Czech Republic. Climatic Change, 1-16. doi: 10.1007/s10584-018-2353-5

Sievers, U. 2016: Das kleinskalige Strömungsmodell MUKLIMO 3 Teil 2: Thermodynamische Erweiterungen. Berichte des Deutschen Wetterdienstes 248.

Žuvela-Aloise, M., Andre, K., Schwaiger, H., Bird, D. N., & Gallaun, H. 2017: Modelling reduction of urban heat load in Vienna by modifying surface properties of roofs. Theoretical and Applied Climatology, 131, 1005-1018. doi:10.1007/s00704-016-2024-2

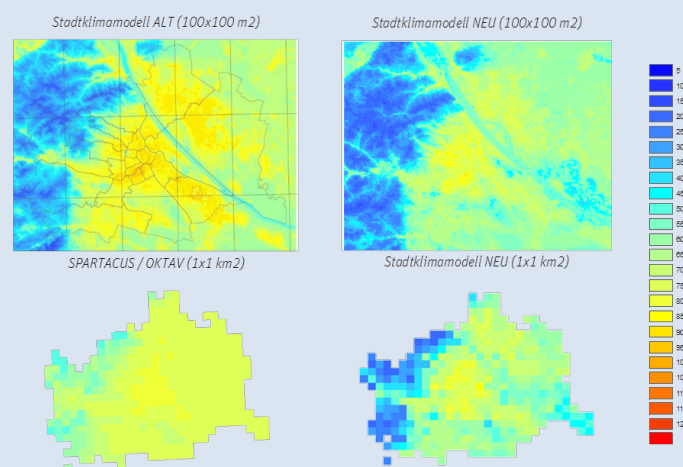


Figure 14 : Comparison of “Downscaling of climate scenarios using urban climate model 100x100 indices” with SPARTACUS/OKTAV climate scenarios for Austria

Urban resilience to extreme weather pilot Mean annual number of summer days for the time period 2021-2050 for Vienna for the scenario c. Comparison of urban climate model (Stadtklimamodell) on 100 x 100 m² resolution, upscaled results on 1 x 1 km² grid, and national SPARTACUS/OKTAV dataset on 1 x 1 km² resolution based on statistically downscaled EURO-CORDEX data.

4.3.5 Provenance Metadata Description

All data transformation has to be described to have a critical understanding of the resulting information, be able to understand the potential biases, assess the skills of the transformations, build trust. This is more and more critical as the users develop more and more expertise in Earth Observation, as they use these products for decisions that are impacting for the society or their business, as fake news and lobbying develop distrust of science.

Box 4-13: Data quality. Lessons learnt from FMI, DWD and ZAMG across several Pilots.

FMI : Use of Statistical methods to adjust SEAS5 Seasonal and ERF Sub seasonal forecasts over Finland

The forecast data used by FMI in the development and production of seasonal climate outlook for the City of Helsinki is provided by the SEAS5 seasonal forecast system of ECMWF accessed from C3S Climate Data Store, where it is updated on the 13th of every month. The spatial resolution of the data is 1° x 1°, the model system includes 51 ensemble members for real-time forecasts. Re-forecast data for the period 1993-2016 was used in the post-processing of the variables. ERA5 re-analysis data accessed from C3S Climate Data Store was used in the skill assessment and bias adjustment of snow data. ERA5 re-analysis was used for the same period as re-forecasts. The variables used are the snow depth and snow density, which following the quality assessment is bias-corrected to reduce the systematic biases from the raw model. The bias adjustment technique applied is the ensemble model output statistics (EMOS) method.

The forecast data used in the development of sub-seasonal climate outlooks are provided by the ensemble prediction system (EPS) of the European Centre for Medium-Range Weather Forecasts. Since the ERF data is not available in the C3S Climate Data Store, it was accessed from ECMWF through the Meteorological Archival and Retrieval System (MARS) in the development phase and through the ECMWF dissemination in the operational service. EPS includes 51 ensemble members, and the forecasts are extended up to 46 days twice a week, on Mondays and Thursdays. The horizontal resolution of the forecast is 0.4° (~36 km) and the surface-based data is available 6-hourly. The variables used are the 2m temperature, snow depth, snow density, and total precipitation. Verification and calibration of variables are done using observations for the period 2000-2019. Observations from Helsinki Kaisaniemi were used in the first year, and, after the discussion with the end users, observations for the Helsinki Kumpula for the second year. Bias adjustment of variables was done by comparing the 20-year model climatology constructed from the re-forecast data to observations. The 20-year model climatology is created not only from the re-forecasts of the same day and month but using a one-week window of re-forecasts.

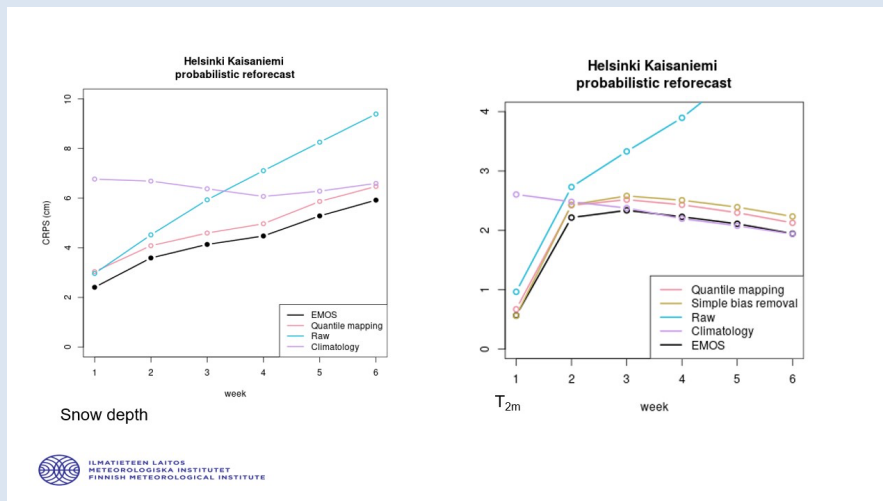


Figure 15 : Snow Depth and T 2m Reforecast adjusted with statistical methods

After testing different methods and taking into account the practical considerations, the temperature was adjusted by removing the average difference between observations and the model climatology from the raw model forecasts before using it in the service. Snow depth forecasts are adjusted using the empirical quantile mapping (EQM) method.

DWD: Downscaling data without losing skills

Climate prediction model

The German Climate Forecast System Version 2.1 (GCFS2.0, Fröhlich et al. 2020) provides the operational seasonal climate predictions for the DWD climate service for the capitals of the German federal states and for Aschaffenburg (e-shape pilot city). It is based on the global coupled earth system model of the Max Planck Institute for Meteorology (MPI-ESM-HR, Müller et al. 2018, Mauritsen et al. 2018), which has a spatial resolution of ~100 km. The German Climate Forecast System is running on the European Centre for Medium-Range Weather Forecasts (ECMWF) high-performance computing facility in Bologna, Italy (Atos BullSequana XH2000). The seasonal climate predictions are updated on a monthly basis and give a prognosis on the development of temperature and precipitation for the coming 1-6 months. To get the best estimate of the initial conditions for the seasonal predictions, observations are assimilated into the MPI-ESM-HR prior to the start of the predictions. The current GCFS2.1 uses continuous nudging to bring the state of the model climate close to the observation-based climate state (Baehr et al. 2015, Fröhlich et al. 2020). For a robust statistical estimate of the quality and reliability of the predictions, a large number of historical forecasts (also called hindcasts) are calculated for each forecast. Each forecast starts with slightly varying initial conditions of the climate system. The resulting variety of solutions, also called ensemble, is used to evaluate the uncertainties caused by the non-linearity of the climate system. In GCFS2.1 the ensemble is generated by different methods in atmosphere and ocean. For the hindcasts, GCFS2.1 starts 30 ensemble members for each calendar month in the years between 1991-present. Real-time forecasts are performed with 50 ensemble members per forecast.

Baehr J., et al., 2015: The prediction of surface temperature in the new seasonal prediction system based on the MPI-ESM coupled climate model, *Climate Dynamics*, 44, 2723–2735.

Fröhlich, K., et al., 2020: The German Climate Forecast System: GCFS, *Earth and Space Science Open Archive*. <https://doi.org/10.1002/essoar.10502582.2>

Mauritsen, T., et al., 2018: Developments in the MPI-M earth system model version 1:2 (MPI-ESM1.2) and its response to increasing CO₂, *Journal of Advances in Modeling Earth Systems*, 11(4), 998-1038.

Müller, W. A., et al., 2018: A Higher-resolution Version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR), *Journal of Advances in Modelling Earth Systems*, 10(7), 1383-1413.

Statistical downscaling

For Germany the spatial resolution of the global climate predictions is increased using the empirical statistical downscaling method EPISODES (Kreienkamp et al. 2019). The increase in resolution is achieved using statistical relationships between large-scale processes (e.g. air pressure) prescribed by the global climate model and small-scale target variables (e.g. precipitation), build up on observations. The 5 km gridded observation data HYRAS (Razafimaharo et al. 2020, Rauthe et al. 2013) is used for precipitation downscaling. Grids of monthly averaged 2 m air temperature over Germany with a spatial resolution of 1 km are used for temperature downscaling. The data is freely available at the DWD Climate Data Centre (CDC). The EPISODES output has a horizontal grid point distance of about 5 km. The seasonal prediction for each city is an average of all grid boxes within the city or county boundary. In these cases, aggregation can sometimes take place using less than the usual minimum 9 grid boxes, as the outputs from EPISODES are based on observation data.

Kreienkamp, F., et al., 2019: Evaluation of the empirical-statistical downscaling method EPISODES, *Climate Dynamics*, 52, 991-1026. <https://doi.org/10.1007/s00382-018-4276-2>

Rauthe, M., et al., 2013: A Central European precipitation climatology – Part I: Generation and validation of a high-resolution gridded daily data set (HYRAS), *Meteorologische Zeitschrift*, 22(3), 235-256.

Razafimaharo, C., et al., 2020: New high-resolution gridded dataset of daily mean, minimum, and maximum temperature and relative humidity for Central Europe (HYRAS), *Theoretical and Applied Climatology*, 142, 1531-1553.

Skill

The quality (i.e. skill) of the climate prediction is evaluated against the skill of the reference prediction climate mean (climatology) observed over the evaluation period. The evaluation period for the seasonal climate predictions is 1991-2020. The skill score of the ensemble mean prediction is determined using the skill score of the mean-squared error between the ensemble mean of the hindcasts and the actual observation (MSESS). The prediction skill score of the probabilistic predictions is determined using the ranked probability skill score (RPSS), which examines to what extent the predicted probabilities of occurrence of the categories (e.g. 'below normal', 'normal', 'above normal') agree with the category actually observed. The reference prediction for the skill score is the observed climatology. A bootstrapping method is then applied to verify whether the skill comparison between hindcasts and a reference prediction is subject to accidental variations due to small sample sizes (significance test). The described data post-processing (statistical downscaling, skill analysis etc.) is done on DWDs high-performance computing infrastructure.

ZAMG: Input Data Provenance

Urban Atlas - The Land Use (LU) classification of the Urban Atlas (UA) was merged with information obtained from the local municipal authority, including nearby districts, to statistically analyse the LU

characteristics. These classification were used to characterize each LU class's basic urban features such as the fraction of buildings, streets, vegetation and bare soil (see below).

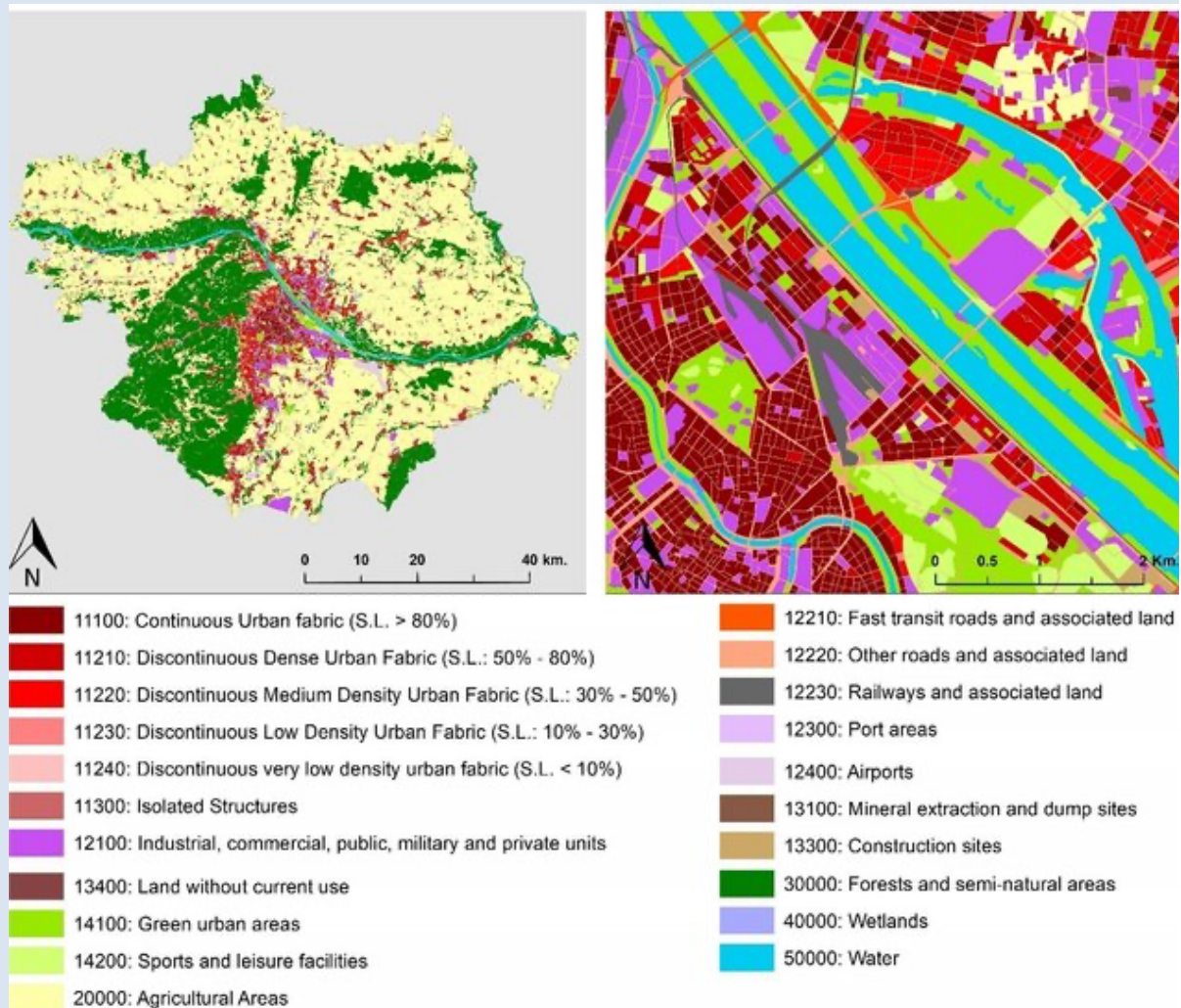


Figure 16 : Land Use refinement with local municipal authority data

LISA - In addition, data from the Land Information System Austria ([LISA](#)) were used, which covered large part of Austria with a 1 m resolution. LISA provides extensive land cover data derived from satellite pictures from 2014 to 2016 and includes eleven distinct land cover types, such as buildings, streets, trees, annual crops, and cobblestone sidewalks.

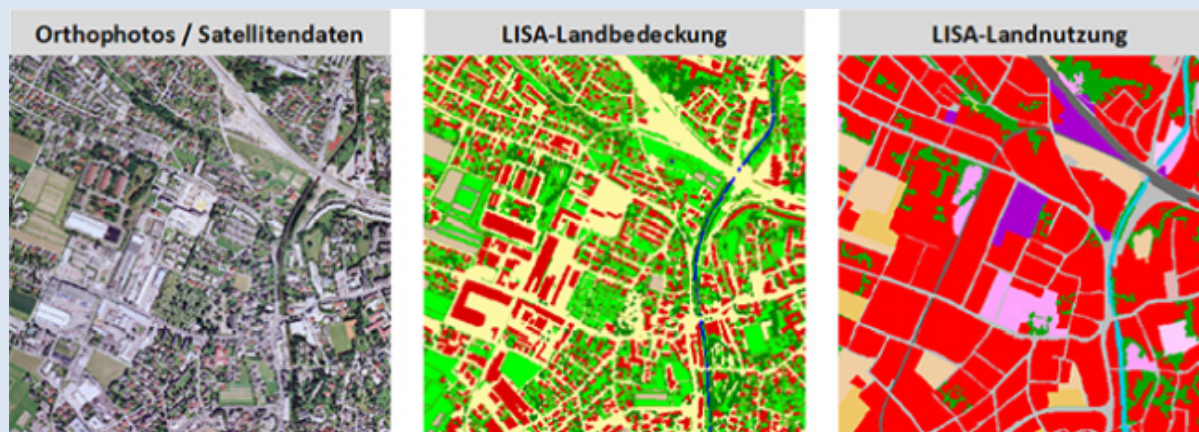


Figure 17 : Land Use refinement with Land Information System Austria (LISA) 1m data

EURO-CORDEX - The World Climate Research Programme launched the Coordinated Regional Downscaling Experiment (CORDEX) with the goal of supporting, coordinating, and improving regional climate scenarios through global collaboration. The EURO-CORDEX research project for Europe aggregated future climate forecasts through Regional Climate Models (RCMs) at 50 and 12.5km spatial resolution based on RCPs as established in the Intergovernmental Panel on Climate Change's Fifth Assessment Report. These models give data on key meteorological characteristics through 2100 under various climate change scenarios.

The pilot used model outputs from three different RCMs combined with six Global Climate Models at the 12.5km spatial resolution under RCP4.5 and RCP8.5 for the time period 2011-2100 to estimate possible future urban climate scenarios from the EURO-CORDEX model database. RCP4.5 is a scenario in which CO₂ emissions peak by 2040, whereas RCP8.5 represents a more extreme scenario in which CO₂ emissions continue to climb until 2100.

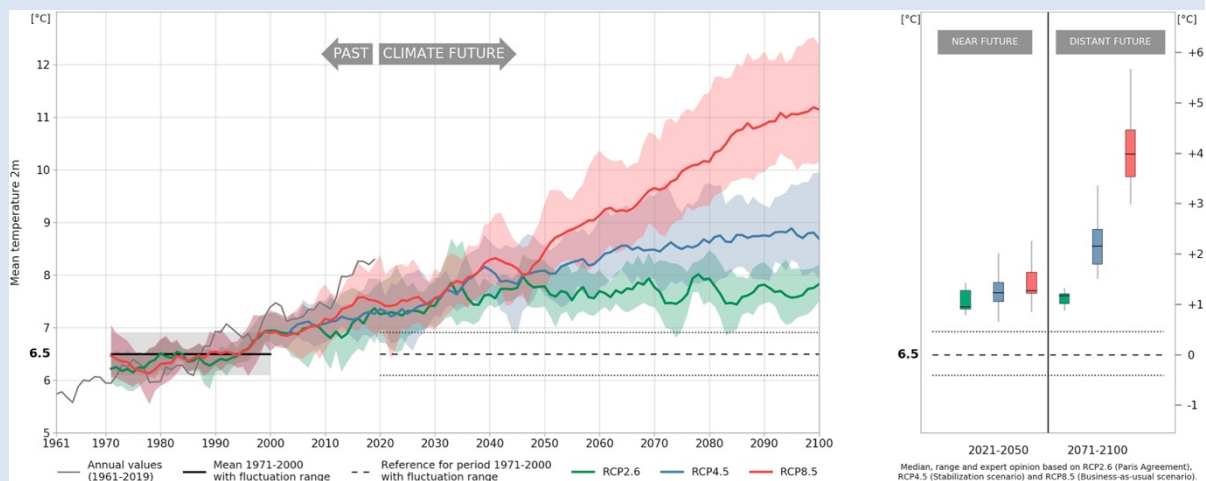


Figure 18 : Past observed (1961–2019) and future projected (5-year running mean regional climate model simulations for scenarios RCP2.6, RCP4.5, and RCP8.5 in the period 1970–2100) annual mean temperatures for Austria (left) and climate change signal compared to the 1971–2000 period (right).

Shaded hues represent the bandwidth per scenario provided by the various climate models, whereas solid lines represent the model median. Source: [Olefs et al. \(2021\)](#)

Olefs, M., Formayer, H., Gobiet, A., Marke, T., Schöner, W., & Revesz, M. (2021). Past and future changes of the Austrian climate–Importance for tourism. *Journal of Outdoor Recreation and Tourism*, 34, 100395.

4.3.6 Gap fill based on AI or Deep Learning

There are several potential approaches to gap-fill data using artificial intelligence (AI) or deep learning (DL). Some common methods are:

- Statistical Time series analysis method for analysing and forecasting time-based data. Machine learning models can be used to fill gaps in time-based data.



-
- Linear regression is a statistical method for modelling the relationship between a dependent variable and one or more independent variables. This approach can be used to predict missing values based on the relationship between other variables.
 - Deep autoencoders are type of neural network that can learn to compress and decompress data. In this approach, deep autoencoders are trained on the available data, and then used to generate predictions for the missing values.
 - Generative Adversarial Networks are a type of neural network that can generate new data that is similar to a given dataset. GANs can be trained on the available data and then used to generate predictions for the missing values.

5 DATA SOURCE(S) SELECTION

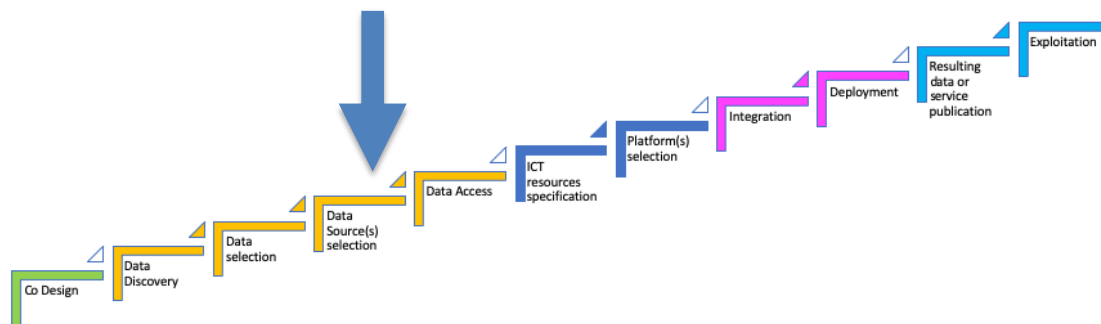


Figure 19: Data source Selection in the Development workflow

5.1 Introduction

The same data sets with the same or different coverage and resolutions in space and time can be accessible from different sources. The temporal depth of the online or offline archive can be different. They can be served in different formats and with different APIs. Data platforms can offer different Service Level Agreements -SLA- for the delivery service in terms of performance, availability, and response time for technical support. For instance, an SLA between a web hosting company and a client might specify the level of uptime (i.e., the amount of time the client's website will be available online) and the response time for technical support requests. When this is business critical, the SLA might also specify penalties, such as service credits or refunds, if the hosting company fails to meet the agreed-upon level of uptime. The following section reviews some criteria that the e-shape pilots have been looking at, implemented, or used.

5.2 Data Cubes Data organization and services as an enabler for efficient exploration of multidimensional data

Data can be accessed from Data Cubes to make multidimensions exploration easier and more efficient.

The OGC® Discussion Paper "OGC: Towards Data Cube Interoperability" defines a data cube as *"a discretized model of the earth that offers estimated values of certain variables for each partition of the Earth's surface called a cell. A data cube instance may provide data for the whole Earth or a subset thereof. Ideally, a data cube is dense (i.e., does not include empty cells) with regular cell distance for its spatial and temporal dimensions. A data cube describes its basic structure, i.e., its spatial and temporal characteristics, and its supported variables (also known as 'properties'), as metadata. It is further defined by a set of functions. These functions describe the available discovery, access, view, analytical, and processing methods that are supported to interact with the data cube."*

It is a 1 to n dimension. Data cubes are often implemented by platforms serving data as a means to facilitate multidimensional exploration and offer good performance over all of them. The meteorology community has used Data cubes to serve numerical models' outputs in the 90ies. Numerical Model data cubes are not always dense nor regular in space and time: there are more layers near the ground than in the upper atmosphere, and more time steps in short range than in longer ranges. Some products are only processed on the levels where they can occur, such as the risk of turbulence or icing. In this case, data cubes fit the production process and are not as regular and dense as the user would expect. The concept has been adopted by the Space community to address the need to explore time series as well

as spatial data exploration. In this case, the data cubes allow to go over the production process and prepare the data access in a more user-centric way than just spatial product collections for a given time that stick to the production process.

As the Data cube concept is a means to bridge the production process specificities to easy and efficient data access for the user, there can be several implementations depending on the data they served and on/or the user community they target. The OGC® Discussion Paper "OGC: Towards Data Cube Interoperability" provides a figure to illustrate several existing options.

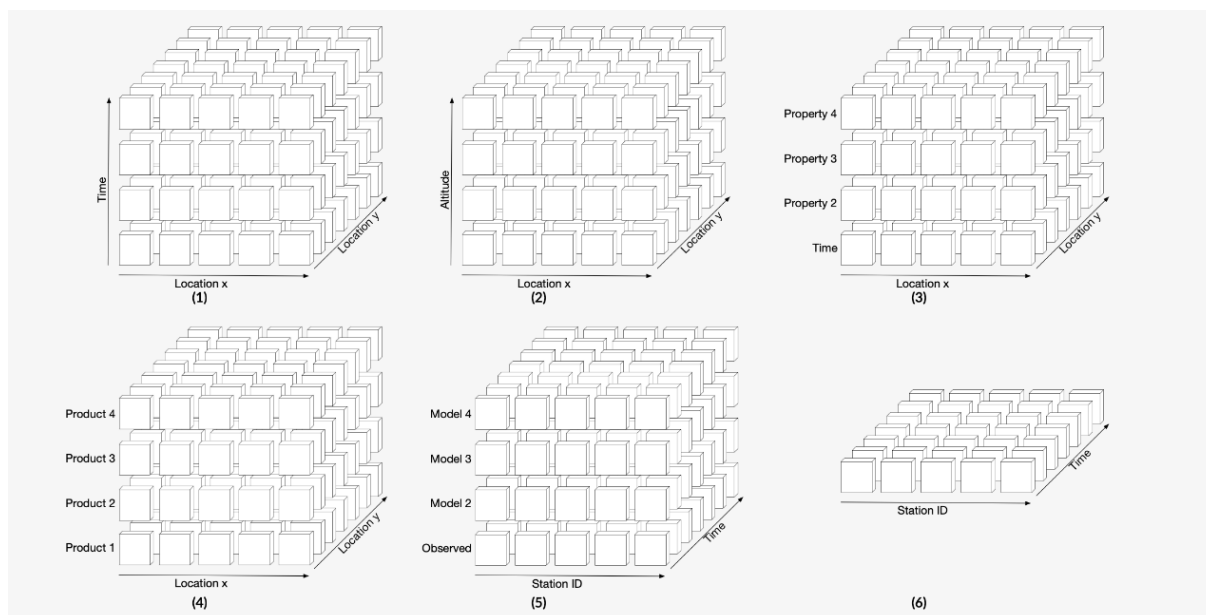


Figure 20: Different options for Data Cubes implementations

The OGC Discussion paper is based on a workshop co-organized by GEO and OGC that gathered data providers and data users. The workshop introduced different types of data cubes organizations as the data providers suggested data cubes homogeneous in terms of data types when the users implemented their own data cubes merging heterogeneous data types but optimized to serve their decision process. This could be mapped with an architecture of Analysis Ready Data Cubes used to build Decision Ready Data cubes.

Box 5-1: Data Cubes. Some Data cubes used by e-shape pilots.

EuroDataCube (<https://eurodatacube.com/>)

The EuroData Cube homogenizes the access to Data from different sources to make them available in one place (one-stop-shop for EO). Combined with the Cloud capacities provided by PaaS Platforms such as DIAS or AWS, it allows processing near the data with a unified API and standards. It is Cloud-agnostic and currently operational at AWS, CreoDIAS, and Mundi. It allows adding custom data by implementing the "Bring Your Own Data" concept including non EO Data via xcube or to generate "data on demand". It implements several OGC standards such as OGC WMS and EO-WMS profile, WCS and EO-WCS profile. It provides an environment to run Jupyter Notebooks and to host your own applications, and a Marketplace for free or revenue-generating options to share data, applications, and algorithms via github enabling cooperation (<https://github.com/eurodatacube>). These are some of the major capacities, please check the website for updated and comprehensive capacities.

Smartmet Data Cube

SmartMet Server is a data and product server for MetOcean data based on data cubes concept, developed by the Finnish Meteorological Institute (FMI). It provides a high capacity and high availability data and product server for MetOcean data. The main feature it carries is being able to return any grid product (same as raster) in another grid with a different resolution/projection. Data from different producers and variables can always be computed in a common grid. Visualization is performed on-demand into images or OGC (tiled) web maps. Gridded data can also be served as OGC web features or downloaded into NetCDF or Grib files. The most powerful interface is the time series plugin. It is a way to query and compute on demand to a point, route, or area several variables allowing equations between them. With lua function, even more complicated processing can be performed. The Pilot 3 of the Climate showcase <https://harvesterseasons.com/> is using this feature to compute from 51 values of a seasonal forecast with 3 variables an index of values 0, 1, or 2. It seeks if 90% of the ensemble members are above a certain threshold for a 2. 0 is the case if 90% of values for this variable are below another threshold. 1 is for the case that good 2 or bad 0 are not reached. See the codes at <https://github.com/fmidev/harvesterseasons-site> as a single request to SmartMet-server time-series API. The API is described at <https://github.com/fmidev/smartmet-plugin-timeseries> with usage examples and the API syntax. The server data in this pilot is also hosting Sentinel-3 NDVI data in addition to bias-adjusted C3S seasonal forecasts ECBSF and ERA5 reanalysis or FMI weather forecasts and postprocessing model data. The data available can be discovered with the grid-GUI plugin at <https://sm.harvesterseasons.com/grid-gui>. The SmartMet server system is cluster ready to enable the latency low enough for web users to be ready to wait for. Splitting the data backends to end-user serving frontends, even large data computations can be served in a short time. The pilot service is not needing this functionality but is still performing with only some latency for the data queries. Powerful caching is enabled to reuse existing queries and speed up the service.

MEEO Data Cube

The ongoing e-shape FRIEND S6P5 pilot (<https://e-shape.eu/index.php/showcases/pilot-6-5-friend>), developed in the context of the climate security domain, relies on the ADAM - Advanced geospatial Data Management technology (<https://adamplatform.eu/>). This technology fully implements the data cube concept, with a pixel based access to multi-dimensional data that are distributed on different data cubes. In particular, the ADAM platform Data Access Service (DAS) represents the component in charge of applying the data cube approach: its implementation in front of each data source enables effective access services.

The FRIEND service is accessible through a web platform which allows the user to assess the flood risk (by mainly exploiting CMEMS, GloFAS, IMERG and ERA5 Land data as risk products), as well as to assess

the flood event impact (through the GEO-DAMP products), for three pre-selected areas (Char-Piya, Australia and Darfur) with a spatial resolution of 10 m. Mainly data related to relevant historical events are considered, but the platform provides also near real time data for the risk assessment. Thanks to the ADAM data cube technology, the FRIEND pilot implements the Digital Earth concept and the data are stored on distributed systems available via standardized Open Geospatial Consortium (OGC)-compliant interfaces.

The FRIEND pilot, in particular, is based on two data cubes (both based in MEE0-ADAM technology): a MEE0 customized DAS (ADAM DAS), for the risk assessment products which are pre-processed and ingested by MEE0, and a SatCen GEO-DAMP DAS which provides access to flood and water masks generated from Sentinel-1 and Sentinel-2 imageries. The ADAM technology offers WMS/WMTS services for all FRIEND raster datasets, providing subset maps that can be interactively navigated, and producing time series over a selected location, while the Web Coverage Service (WCS) is the core part of the DAS module. WCS can be queried directly via REST queries. The data cube technology is used by the FRIEND service for all the crucial operations in terms of discovery, visualization of maps and time series, processing and download of the data related to the selected specific location.

Lessons learned on Geospatial Data Cubes

The recourse to a Data cube service can be an efficient strategy when selecting a data source, it is however probable that a benchmarking exercise should be conducted to identify the most effective source for a given requirement scenario.

Data Cubes optimize the access to multidimensional data but they can be implemented with differences in design, interfaces or dimensions characteristics leading to interoperability issues when there is a need to interact with several data cubes.

References on Data Cubes

- Euro Data Cube: <https://eurodatacube.com/documentation/about>
- Simonis, I. 2021:OGC® Towards Data Cube Interoperability 21-067
- OGC initiative Data Cube Interoperability: <https://www.ogc.org/projects/initiatives/gdc>
- GitHub page for using SmartMet-server: <https://github.com/fmidev/chile-smartmet/>

5.3 Umbrella Sentinel Access Point (USAP) as an enabler for data access efficiency and resilience

NOA has developed the so-called Umbrella Sentinel Hub (USH), which acts as a broker among multiple Sentinel Access points (i.e. Open Access Hub, Hellenic Mirror Site, Finish Mirror Site), Sentinel 5P hub etc.). This application allows the user to access data from all Sentinel missions through a single API. Additionally, continuously harvesting metadata from multiple hubs allows for reduced latency and increased download speed (through the USH's scoring mechanism of the connected Hubs). It should be noted that USH accesses only Sentinel data and no other data from GEOS and GEO.

USH allows for the seamless access of Sentinel data through a single access point. This provides complete global coverage, reduced download speeds, reduced latency of ingestion and complete availability to all Sentinel missions. USH follows the same API logic, as the DHuS based Sentinel Hubs (e.g. Open Access Hub) to allow for the seamless transition of other Sentinel data access scripts and applications from the existing Hubs to USH.

USH is already deployed and operating on NOA owned infrastructure and can be easily deployed on other cloud environments. Nonetheless, there is no particular interest on where exactly USH is deployed, as it is a broker of metadata and not a database that physically stores the data"

Box 5-2: A single point of access for sentinel data, connecting multiple data sources.

Searching for Sentinel data is often a complicated process due to the different missions available and the different hubs that host the data, but also to the different performances of the hubs in terms of download speed and latency (at both the inter and intra level).

Thus, the Operational Unit BEYOND of IAASARS/NOA has developed the Umbrella Sentinel Access Point that brings them all together (Umbrella Sentinel Access Point (beyond-eocenter.eu).

- The Umbrella Sentinel Access Point acts as a single data access point which:
- links Sentinel data hubs, regardless their back-end architecture, to a single data hub provides access to all Sentinel mission data and better performance on downloading products, as products are chosen from the most appropriate data hub at any time instance based on integrity, speed and availability tests.

Currently, the Umbrella is being extended in order to allow the potential users to blend also meteorological parameters, such as temperature, aiming at decreasing the number of satellite images provided based on this parameter.

The Umbrella provides the aforementioned services via an API, an example of which is presented below:

http://umbrella.beyond-eocenter.eu/api/products/sentinel2?in_bbox=20.8,38.41,23.82,40&sensing_date_gte=2020-05-01&sensing_date_lte=2020-05-31

5.4 General considerations

If an application uses data from one single data producer, getting the data directly from the data producer in a short circuit will be efficient and reliable.

If an application uses data from several data producers, getting the data directly from the data producers will make the data retrieval complicated, going through a data hub will homogenize the access. (ex: DIASS, NextGEOSS)

If an application uses data that is duplicated and made available on several platforms, a Data Broker can provide more reliable and efficient access to complete global coverage of data. (ex Umbrella Sentinel Hub). This module can be available on several platforms (GAIASENSE, NextGEOSS...)

6 DATA ACCESS

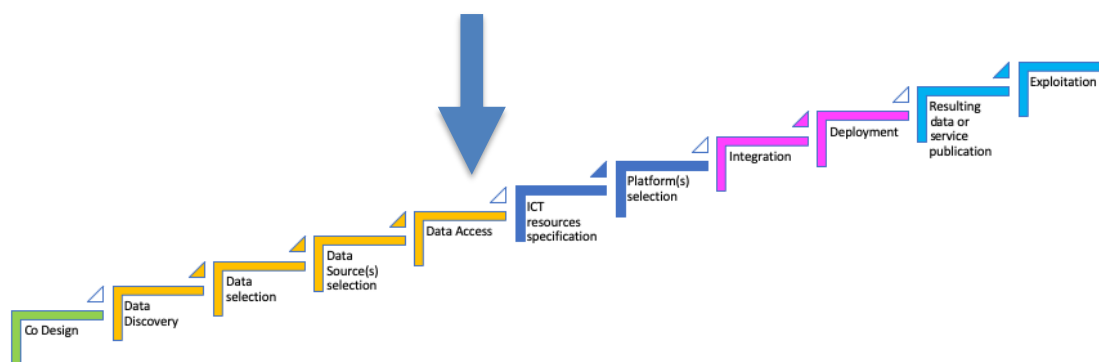


Figure 21: Data access in the Development workflow

6.1 Introduction

Data access can be challenging in various contexts due to a range of factors, including data privacy and security, technical limitations, legal and regulatory requirements, data quality and reliability, organizational and cultural barriers, cost, and data complexity. Despite these challenges, data repositories provide value by making data discoverable and enabling analysis, knowledge extraction, and dissemination. The rise of open data and open science data trends, along with the implementation of data management and sharing policies by funding bodies, governments, and private companies, has led to an increase in research data services. Access to publicly accessible data can benefit research reproducibility, reliability, visibility, and future research. It can also attract new partnerships and allow for the use of new innovative research approaches and tools. Last but not least, it enables the development of new businesses and new services bringing societal value. Access strategies should be defined for each specific situation, taking into account seamless access and resource combination. It supports the business model as the access strategy can vary depending on the audience (research or business), over time (open during a project development then restricted or restricted in real time then open when the data gets older), on the resolution (free at low resolution, paying at high resolution), free for some samples and paying for regular updates...

6.2 Data Privacy and security

When dealing with Citizen observations or in situ in Europe, the General Data Protection Regulation (GDPR) has to be addressed. Other countries also have their own regulations. GDPR is a regulation of the EU legislation on data protection and privacy in the European Union and the European Economic Area. It also addresses the transfer of personal data outside the EU and EEA areas. The GDPR's primary aim is to give control to individuals over their personal data and to simplify the regulatory environment for international business by unifying regulations within the EU.

The French National Institute for Information & Rights Protection (Commission Nationale Informatique et Liberté - CNIL) provides a toolkit to support compliance to GDPR. (<https://www.cnil.fr/en/gdpr-toolkit>) that includes a tool to carry out a Privacy Impact Assessment -PIA (<https://www.cnil.fr/en/privacy-impact-assessment-pia>)

When working with personal data outside of Europe these considerations have to be reviewed and adapted to local regulations. This approach will typically differ by country and depend on current legislation and data-sharing policies: e.g. Tanzania is extremely strict, some organizations elsewhere require signed contracts and/or payments, while most nations/organizations require basically the presence of local staff (approved presence) during fieldwork and interviews.

Besides the ethical considerations, geolocation data can provide valuable information about the location and movement of vessels or planes, which can be exploited by malicious actors for various purposes, such as theft, piracy or terrorism. Earth Observation is geolocated data and therefore, this has to be handled carefully.

Box 6-1: Data privacy & security. Lessons learnt from Showcase 1: [Food Security and Sustainable Agriculture](#) Pilot3: [Vegetation-Index Crop-Insurance in Ethiopia](#) (drought insurance for smallholder farmers) Pilot

In the Agriculture VICI-Ethiopia Pilot, GDPR management is basically catered for through the established Public-Private-Partnership architecture. Locally, the role of ITC is limited to academic work/advice regarding the use and quality of remotely sensed imagery to capture impacts by perils (mainly of droughts) and calibration of VICI thresholds. Implementation is fully carried out through Ethiopian public & private partners of the consortium. Any action/intervention/training/sale that is carried out in Ethiopia always takes place "through" or at least "with" (under the responsibilities of) one of our local PPP members. These national partners guarantee that we (ITC) properly adhere to the local GDPR aspects. In 2021, working locally on VICI validation and perils-inquiries will mainly be carried out through the ICIP project that has in turn direct connections with (oversight by) MoAgr, ATA, and Kifiya. In short: we work with local partners who locally fully carry out all responsibilities regarding local regulations etc.

Box 6-2: Data privacy & security. Lessons learnt from the [Water resources management](#) showcase, the development of Pilot 5 [Monitoring fishing activity](#)

Access to datasets with in-situ recordings of commercial activities with a spatio-temporal scope are regarded as fundamental and complementary to remote sensing data in a broad range of application studies. Moreover, the EU DIRECTIVE 2019/1024 on open data and the re-use of public sector information acknowledges that the EU open data market is a key building block of the overall EU data economy.

Currently, data collection and their timely access have increasingly been regarded as key factors for a sustainable exploitation of marine resources and for better environmental protection of the oceans.

Within the [Water resources management](#) showcase, the development of Pilot 5 [Monitoring fishing activity](#) was extremely challenging due to recurrent difficulties concerning full access to the required data about the activity of fishing vessels. The National Fisheries Authority was enrolled in this pilot as a data provider, ensuring access to part of the fisheries-dependent monitoring data from the activities of national vessels, including fleet characteristics, electronic logbooks, and landing datasets, with anonymized identification of the vessels. However, VMS (vessel monitoring system) data which accounts for the spatial behaviour of fishing vessels activity, collected through the MONICAP system, have not been provided for the Pilot. Aware of the relevance of these data for the pilot's outcomes, it has been partly circumvented by acquiring AIS (sat-AIS) data from external suppliers; however, insufficient coverage of satellite data, together with the specific nature of this system, led to gaps in fisheries trajectories, thus affecting the quality and usefulness of the final products.

On the other hand, fisheries-dependent monitoring data for other EU and foreign vessels fishing in the study area were unavailable to be shared at a national level. Despite efforts from e-Shape to support the release of these data through formal contact with the international organizations managing fisheries in the Northeast Atlantic, this information could not be obtained.

From a development and operational point of view, this difficulty in accessing fisheries monitoring data hinders the accuracy and scalability of the application and was therefore one of the main obstacles to the development work done by the pilot team.

Lessons learned on Data privacy

- Data Privacy is a legal requirement which needs to be complied with. It is an integral part of the sustainability plan of an application (whether commercial or open source)
- Data Privacy is a sensitive issue that has to be addressed properly as soon as possible to be sure of the usability of the datasets
- It can require the support from an expert, hopefully local to the country where the measurements are done

References on Data privacy:

- Complete guide to GDPR compliance <https://gdpr.eu/>
- CNIL's GDPR Compliance toolkit: <https://www.cnil.fr/en/gdpr-toolkit>
- CNIL's Privacy Impact Assessment Software: <https://www.cnil.fr/en/gdpr-toolkit>

6.3 Technical limitations

Many standards have been developed and adopted largely by the community to enable online access to frequently updated data. Currently, the access is often enabled via a URL responding to HTTP(S), or secured variants of FTP based protocols, standardized web services via APIs such as OGC APIs family (Environmental Data Retrieval, OGC API - Features, OGC API - Coverages, Sensor Things API) or web services such as OGC Web Map Services / Web Map Tiles Services, OGC Web Coverage Services... and more a more also direct processing on the cloud.

Nevertheless, diverse technical challenges remain to be addressed:

- First are the legal aspects, the data licenses, and access control. If most of the e-shape pilots have used Copernicus Open data, very few of them have only used open data and EO developments build most of the time on data with different access restrictions that can be handled by buying data when the data is downloaded and/or with authentication services when it is accessed online.
- Accessing data distributed over several locations can involve Single Sign On solutions where the user will identify once and be recognized by several systems. When the data is really fragmented Implementing a centralized data management strategy, such as a data lake or a data spaces, can ensure that data is stored in a single location, homogenised, (hopefully) curated, making it easier to access and manage. The concept of thematic Data spaces is currently developing rapidly in Europe, boosted by the increasing amount of IoT data and all the businesses easy access to data can enable.
- Network bandwidth, latency, or reliability can also impact connectivity and affect the speed and quality of data access, particularly for large datasets or real-time applications. Reducing the amount of data to be transferred by using filters, Standards implementing tiles, processing the data in the cloud to transfer the value-added information only can address this issue. Other standards such as OGC GeoPackage Standard can address this, for instance, GeoPackage is based on SQLite local database file format that stores, efficiently storing, sharing, and managing various types of geospatial data, including vector and raster data, in a single file format.
- Considering Data formats, standards again are widely adopted to facilitate the data access but the variety of data that can be vector, raster, regular in time or event-driven, 1 to many dimensions involves many standards that can require converters, parsers, or schema definitions. OGC vision is to use location to facilitate the filtering and as a driver to interoperability over data silos and data heterogeneity.

- Data volume can be a technical limitation to data access, particularly for large datasets or high-throughput applications. Earth Observation is a typical example of big data and accessing large datasets may require specific technologies, such as distributed computing, parallel processing, or data sharding.
- Data security requirements may limit data access by enforcing encryption, anonymization, or masking of sensitive data. Accessing encrypted or anonymized data may require specific tools or technologies, such as encryption keys, decryption algorithms, or anonymization protocols. This can have to be considered for Data privacy considerations for instance.

As seen above, Data Access issues are addressed via standards, and technical architectures involving data centralization in different data management systems such as Databases, Repositories, Data Spaces.

[Agrostack](https://agrostack.org/en) (<https://agrostack.org/en>) developed, for instance, by the e-shape Agriculture Showcase makes agriculture-related data more accessible. DEIMS-SDR implemented by the Biodiversity Showcase makes Ecological homogenized and curated information related to specific sites of interest accessible.

The Umbrella Sentinel Access Point presented in the chapter on "Data sources selection" is another type of original solution to improve the reliability and performance of the access to sentinel data.

6.4 Legal and regulatory requirements

The European Commission conceived Copernicus' data as full, free, and open to allow the scientific community and developers to use Sentinel data and other Copernicus data without legal restrictions. The goals are to enable science to take full advantage of the value of Copernicus and to foster the development of businesses. By "no legal restrictions," it is meant that users can obtain Sentinel and Copernicus data without paying any fees and can distribute, reproduce, or publish from the source or data provider, which is the European Commission. If most of the e-shape pilots have built their service or products based on Copernicus, they have most of the time also used paying data. The constraints are most of the time lighter for research or development activities but it can be a real issue when building a commercial service or product and it is always better to anticipate. The dynamic to open or share the data keeps on progressing as new datasets are made open regularly.

It is required to make sure that the license of the different components (data, services, software...) used in the production process are clearly understood. This exercise is an integral part of the development of a specific Data Management Plan, in compliance with the GEO and FAIR Data Management Principles, such as developed within the e-shape project.

Reference on compliance:

- e-shape Data Management Plan toolbox: <https://gkhub.earthobservations.org/records/0ksgt-7v316>

6.5 Data complexity

Data complexity can be a real issue. As mentioned earlier in the Analysis Ready Data paragraph, the concept of Analysis Ready Data has been initialized by CEOS to lower the effort and level of expertise needed from the user. ISO, OGC, and CEOS are joining efforts to extend this concept and standardize more Analysis Ready Data products, including for in situ or other EO Data types. This should foster easy reuse of the ARD products and increase the effective value of the Earth Observation data.

References on Data access:

- GEO Data Management Principles Implementation Guidelines: <https://gkhub.earthobservations.org/records/gg85h-x8466>

6.6 User interaction: accessing the Data or running the application

When users want to make a decision, they can access data or pre-processed products, or they might run a service to process the product on the fly via different means.

Two capabilities have to be considered: discovering data, downloading it, and running local processing or discovering applications, uploading them (if they are not already accessible near the data), and processing near the data.

6.7 Login service

Logging services are critical components of modern software systems, providing developers with the ability to track and analyse system behaviour and diagnose problems. Different forms of logging services exist, including system-level, application-level, and cloud-based logging, each with its own specific use cases. Logging services are essential in ensuring the stability and reliability of modern software systems, and they come in various forms, scalability, and customization options. Single Sign-On (SSO) is another form of logging service that enables users to access multiple applications or systems in a federated environment, with a single set of credentials, improving security and streamlining user management. SSO solutions can be deployed on-premises or in the cloud and can integrate with various identity providers and authentication protocols. The SSO Solution is provided by the platform or Cloud provider. NextGEOSS for instance offers a User Management service based on OpenID Connect (OIDC) for authentication and UMA for single-point authorization management.

References:

- NextGEOSS User Management: <https://nextgeoss.eu/wp-content/uploads/User-Management-User-Guide.pdf>
- OGC User Management Interfaces for Earth Observation Services: https://portal.ogc.org/files/?artifact_id=54929

6.8 Earth Observation Data portals

Data portals are web-based platforms that allow users to search, browse, and download EO data. The European Space Agency's (ESA) Sentinel Data Hub is a Data Portal for instance.

If the data products have been pre-processed in batch processing, the pre-processing can be implemented with different strategies: regularly on time conditions (for example every day at 1 am UTC on data from the previous day), it can happen on resources availability conditions (for example: as soon as a satellite product is made available) or it can be event driven (example: when a volcano eruption happens).

6.9 Application Programming Interfaces (APIs)

More and more data providers offer (Application Programming Interfaces (APIs), which are mechanisms that enable two software components to communicate with each other with an agreed set of protocols and definitions for building software applications. Users can use these APIs to access EO data programmatically and integrate them into their own software or workflow.

6.9.1 Mobile applications

Mobile applications providing access to EO data and services, such as satellite imagery and weather information are quite popular. There are many more related to Earth Observation and for instance, pilot 3: [Diver Information on Visibility in Europe](#) (coastal water quality monitoring) of Showcase

5: [Water resources management](#) _ has developed and published an application to provide near real-time visibility in the water score for specific dive locations.

Mobile applications have also been used for data-collection (CropObserve in support to the [GEOGLAM Pilot - The Food Security and Sustainable Agriculture](#) showcase), but also to collect user feedback on the insurance products in Ethiopia by the [Vegetation-Index Crop-Insurance in Ethiopia](#) pilot in the [Food Security and Sustainable Agriculture](#) showcase.

Mobile technologies, can be useful from data collection to final products access and user feedback.

These mobile applications can have diverse business models that will impact the user access to the data such as freemium, one-shot paid app, subscription, in-app purchases, advertising, sponsorship, or commission based.

6.9.2 Online analysis tools and on-demand processing

Due to the increasing volume of data, the more and more frequent data updates, the cost of storage, and the increasing diversity of value-added post-processing, it becomes every day more difficult to pre-process all the possible EO products. This encourages progressively data providers to make a bigger part of their products accessible as on-demand services or via online analysis tools.

Online analysis tools are pre-built software applications that run on servers and allow users to interact with and manipulate data through a web-based interface. The user does not need to download the data, but the tool may need to fetch and download some data to process it on the server. Examples of online analysis tools include the ESA's Sentinel Toolbox and the Google Earth Engine.

On demand processing will trigger a process transforming on-the-fly, raw data or Analysis Ready Data (ARD), into the product required by the user and provide the results via a Web API. Web APIs are a specific type of API that enables the dialog between a Web browser and a Web server. This method can be more flexible and customizable than online analysis tools, as users can define the specific processing tasks to be performed on the data.

In both cases, the user can eventually have to define input parameters to initialize the processing, and this processing can fetch and download the data to process it locally or run a process near the data if the data is too big to be transferred for instance.

6.10 Jupyter Notebook

Jupyter Notebooks are very popular for providing user access to data or applications for training and research. The goal of Project Jupyter "is to build open-source tools and create a community that facilitates scientific research, reproducible and open workflows, education, computational narratives, and data analytics". The Jupyter Notebook is an open-source web application that allows you to create and share documents that include executable code or visual interactive components resulting from the execution of editable programming cells. Jupyter features the ability to execute code changes on the fly from the browser, display the result of computation using visual representation, and/or facilitate user interactions via graphical widgets. It can allow for building dashboards. Jupyter Notebooks are available on all the DIASs, NextGEOSS, ... At least 8 e-shape pilots have used Jupyter Notebooks.

7 NEW PARADIGM OF DIGITAL TWINS

A digital twin is a virtual replica of a physical object or system that is designed, developed, and maintained throughout its lifecycle. It is a digital representation of a real-world object or system, which can include physical products, machines, buildings, or entire cities.

The digital twin is created by combining data from sensors, internet of Things (IoT) devices, and other sources to create a digital model of the object or system. This digital model can be used for a variety of purposes, including design, simulation, optimization, monitoring, and maintenance.

This concept has been gaining traction in recent years due to advancements in digital technologies such as IoT, big data analytics, and cloud computing. The concept of Digital twins applies to diverse domains such as healthcare, manufacturing, construction, energy, and agriculture as a digital copy of the real world.

One of the earliest references to the concept of digital twins in the context of Earth observations can be found in a report published by the European Space Agency (ESA) in 2017, which discussed the potential of digital twins for Earth system modelling and analysis.

Since then, the idea of using digital twins for Earth observations has been included in various European research and innovation programs, such as Horizon 2020, which includes a focus on the development of digital twins for the monitoring and management of natural resources and ecosystems.

Currently, major projects and initiatives are under development to build powerful digital twins to improve the services, reduce costs and optimize the performance, gain real-time insights, identify potential problems before they occur, and make more informed decisions.

Destination Earth (DestinE), is an ambitious initiative of the European Union to create a digital twin – an interactive computer simulation – of our planet. It is currently building a digital twin engine and the first two digital twins:

- The Digital Twin on Weather-Induced and Geophysical Extremes will provide capabilities for the assessment and prediction of environmental extremes in support of risk assessment and management.
- The Digital Twin on Climate Change Adaptation will support the analysis and testing of scenarios. This in turn will support sustainable development, climate adaptation, and mitigation policy-making at multi-decadal timescales, at regional and national levels.

Iliad is another EU-funded project, that builds on the assets resulting from two decades of investments in policies and infrastructures for the blue economy and aims at establishing an interoperable, data-intensive, and cost-effective Digital Twin of the Ocean.

Digital twins were not developed or used by e-shape pilots as such but it looks to be a concept worth introducing here as Numerical Modelling and Artificial Intelligence that contribute to Digital twins have been used by several e-shape pilots and the activities developed during the project will, for sure, be useful to the Digital twins implementation.

Box 7-1: Digital twins. Lessons learnt from [MyEcosystemPilot 2: mySITE](#) (data provision, visualisation tools and ecosystem status indicators)

A number of activities meet the requirements for accessible and interoperable data on the pan-European scale. In this respect, the focus of the EC on setting up a Common European Green Deal data space[1] to support European Green Deal priority actions will play an important role in the further enhancement of data availability and accessibility. This includes initiatives like “GreenData4All” and ‘Destination Earth’ (Digital Twin Earth) as defined by the European Digital Strategy. While the first will look into the review of the INSPIRE Directive (2007/2/EC), the latter shall develop a digital modelling platform to monitor and forecast natural and human activities with respect to sustainable development. The development of high-precision digital models of the Earth System implementation, e.g. Digital Twins will play an important role in supporting policy. Currently, a number of digital twins are under development in the frame of Horizon Europe, e.g. [BioDT \(Digital Twin for Biodiversity\)](#), [DTO \(Digital Twin Ocean\)](#), building data flows and modelling frameworks using AI to address future needs.

The relevance of harmonized data, including Earth Observation as well as in-situ data for environmental and climate assessments, is e.g. summarised in the JRC technical report on “Destination Earth - Use case analysis”[2] focusing on 30 different use cases to support the environmental agenda and policies building on accessible data from various sources including in-situ data.

In this respect, work done in the frame of e-shape can contribute to providing harmonized services on EO data as well as in-situ data to populated models and analytical workflows and build the digital twins for the different environmental sectors addressed. This implies the implementation of FAIR principles to enable machine-actionable metadata and data.

References

[1] See <http://dataspaces.info/common-european-data-spaces/#page-content>

[2] https://publications.jrc.ec.europa.eu/repository/bitstream/JRC122456/destination_earth_usecases_final-approved.pdf see

Lessons learned:

- Open and FAIR provision of data as service is crucial for the implementation of the Digital Twins

e-shape efforts in developing the skills on data management principles and providing a Self-assessment tool to support effective progress that has been endorsed by the GEO secretariat, will be useful to the diverse digital twins implementations.

References on Digital twins:

- Destination Earth:
 - <https://www.ecmwf.int/en/about/what-we-do/environmental-services-and-future-vision/destination-earth>
 - <https://digital-strategy.ec.europa.eu/en/policies/destination-earth>
- Iliad
 - website: <https://www.ocean-twin.eu/> and Iliad
 - Project description: <https://cordis.europa.eu/project/id/101037643>

7.1 New Processing Technologies: Artificial Intelligence, Machine Learning, Deep Learning

Artificial Intelligence has a long history and has evolved in time with the progress of computers and technologies, more intensive use of mathematics, and several periods of hope and disappointment. Literally, it is Intelligence demonstrated by Machines as opposed or as an analogy to Intelligence demonstrated by humans. It is the source of dreams and fears that require to support scientific and technological progress with adequate ethical and legal frameworks as well as socio-economic adaptations. As human natural intelligence, artificial intelligence has to be used with good ethics, morals, and social benevolence and be constrained by some legal framework for a better good. The intelligence being multi-shape, the concept of artificial intelligence is also multi-shape and AI is in fact an umbrella name for different technologies. A common point between these technologies is that the more data and the more multidisciplinary teams, the better. Of course, good quality data makes it even better but we will see that AI allows us to face data lacks and gaps or improve its quality.

In EO, Artificial Intelligence is used at each stage of the value chain where the data is processed: on the satellite itself, in central processing data centres processing Analysis Ready Data, in downstream applications to convert data into information ready for decision. Artificial Intelligence and Digital twins are expected to enable to release the real value from the exponential growth of data. This is why it is considered as a critical technology for further progress. It benefits from massive investment from the public and private stakeholders to move from generating data to generating software that produces information. New fields of application for AI are still to be discovered. The COVID-19 exceptional situation has been the opportunity to identify unexpected uses cases for AI and Earth Observation in connection to health pandemics such as the impacts of the pandemic on human activities, the impacts of the reduction of industrial activity on air pollution, how some conditions like humidity or temperature could favour the spread of that pandemic ...

Machine learning (ML) and Deep learning (DL) techniques are currently the AI-related techniques that are the most used in the Earth Observation Domain. They enable new approaches to cope with the huge increase in available data and benefit from the increase of more affordable computing power and the emergence of efficient algorithms. Machine Learning can be used for many different types of use cases: real-time quality control, data assimilation, numerical modelling, product generation and post processing such as feature detection, error correction, data mining...

Box 7-2: New processing technologies. Lessons learnt from Showcase 1: [Food Security and Sustainable Agriculture](#) Pilot 1: [GEOGLAM](#)

Translating vast amounts of EO data into concrete information can be a huge task, but with the increased computing capacities and data availability, new possibilities are emerging to deploy AI technologies for more performant algorithms. One specific area where AI has proven to be very useful, is to capitalize on the information that is present within a time series, where much of the information is present in how the values change in time rather than the absolute values themselves. In the [GEOGLAM](#) pilot, AI-technologies, and more specifically Sequential Neural Networks, were used to translate Sentinel 1 and Sentinel 2 time series in harvest dates.

While this proved to be a very performant methodology, it was quite dependent on the availability of properly labelled data with sufficient variability in harvest events. This labelling is often identified as one of the main bottlenecks of AI-implementation and upscaling in EO service development. For a proper deployment of this AI-based harvest detection at scale, additional reference data collection and labelling efforts were needed to ensure a proper training of the SNN-model.

7.2 Cloud technologies for Earth Observation

7.2.1 SWOT Analysis of Cloud Technologies for Earth Observations - Theory and Practice

A SWOT analysis is a structured review of a business, project, system or technology emphasizing:

- its Strengths, i.e. characteristics of the technology that gives it an advantage over others;
- its Weaknesses, i.e. characteristics of the technology that places it at a disadvantage relative to others;
- Its Opportunities, i.e. elements that the technology could exploit to its advantage,
- its Threats, i.e. elements that could cause trouble to the technology,

SWOT analyses for cloud technologies are easily findable on the web. We offer here to review here the theoretical SWOT of Cloud technologies for Earth Observations, which can have some specificities and are not available on the web, and the related return of experience from the e-shape pilots introducing a Theory and Practice approach with the lessons learned from e-shape implementation activities. After this Theory and practice review organised with a systematic SWOT approach, the paragraph 7.3 offers a synthesis of the lessons learned for the different profiles of EO Stakeholders.

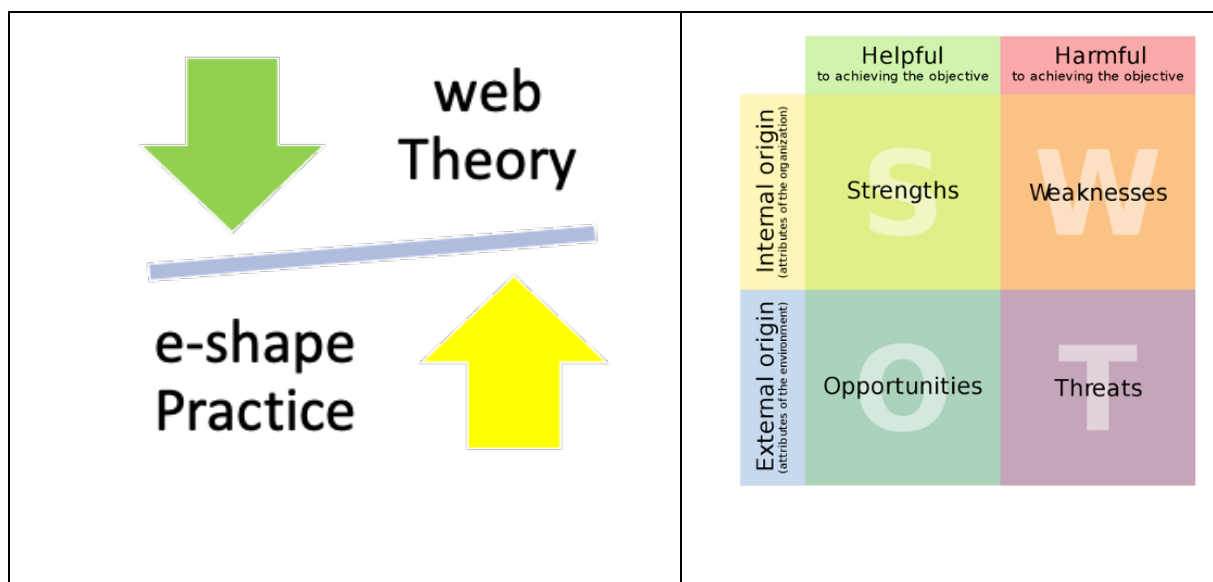


Figure 22: e-shape contribution to a SWOT analysis of Cloud technologies for Earth Observations

7.2.2 Strengths

7.2.2.1 Theory

- driver for Cloud technology's popularity and its fitness to EO.
- Develop a business model as Everything as a service.
- Cost-effectiveness: Reduced expenses based on real use of the resources and lower staff costs in infrastructure management and monitoring
- Flexible and resilient disaster recovery.
- Pricing Transparency by the service providers.
- Faster provisioning of systems and applications

- Secured infrastructure

7.2.2.2 Return of experience from the e-shape pilots

- Compare prices of the different platforms. The temporary free sandbox offered by all platforms can help for this.
- Components from the application or service architecture that can have different needs might be packaged in different containers: for instance, Jupyter Notebooks or front-ends might require more RAM to face randomly increasing users access than CPU or storage when data processing might require more CPU or storage than RAM. Having modular architecture is a general best practice and is very needed in Cloud for EO.
- It is better to minimize the initial configuration as scalability is a native capacity than paying for unused resources
- The price assessment tool from the Network of Resources - NoR can help compare the platform's offers.
- Data as a Service has been the first popular service developed on the Cloud for EO; with the change of paradigm of bringing the process to the Data, Science as a Service or Applications as a Service development is increasing.
- Start-ups will be able to afford infrastructures they cannot buy on their own but major research organizations will keep on using their in house HPCs until a real cost assessment is done, a strategy for the best use of internal/external resources is defined and eventually, a budgeting reorganization is implemented.
- Some pilots had irregular access to the data and had to implement their own data access to secure the service because the support was not reactive enough for the level of reliability their service was demanding
- Pricing is transparent but pricing lists are not always clear and real costs including scalability can become opaque.

7.2.3 Weaknesses

7.2.3.1 Theory

- Specific training required
- Challenge in migrating from one Cloud service provided to another.
- Lack of interoperability between the different cloud service providers.
- Application & Service access is highly dependent on Network Bandwidth.
- Data transfer bottleneck
- Open Standard Implementation.

7.2.3.2 Return of experience from the e-shape pilots

- as a major source of problems. At most some delays that they could mitigate
- No major problems were expressed related to interoperability
- Some pilots have expressed that they were reaching the limits for Data download while others have directly implemented the new paradigm of Applications near the Data that has been designed to mitigate this issue

- Lack of In Situ global or thematic collections push users to develop their own data hubs
- Open standards are used by all the stakeholders from data providers to technology providers, and application developers. Their benefits are obvious including the availability of open-source integration

7.2.4 Opportunities

7.2.4.1 Theory

- on investment in a short time.
- Good opportunity for SMEs to optimize upfront investments,
- Pay-for-Use licenses,
- Adaptive to future needs.
- Cloud provides an excellent backbone for Mobile & Web-based applications.
- High-tech work environment offering modern information solutions according to the last technology,
- Easy, Quick & Low-effective mitigation of identity, privacy, security, reliability, and manageability risks in cloud-based environments.
- The cloud computing approach speeds the deployment while preserving dynamic flexibility.
- EO Platforms provide access to big catalogues of Open Data and Open source
- The EO platforms often offer software packages enabling expert EO data processing.

7.2.4.2 Return of experience from the e-shape pilots

- the adoption of the Cloud technology is a source of complexity and requires developing new skills, involving new skills background in the team or subcontracting part of the activity introducing delays
- the higher return on investment is not clear when the Cloud platforms do not have the same level of "operationality" as the usual resources as debugging or running analysis in a distributed environment can be complex and costly.
- No problems have been reported with identity, security, and manageability risks in cloud-based environments.
- Privacy stays an issue at a different level including the use of Cloud. Reliability has been criticized by several pilots
- The cloud computing approach speeds the deployment for a newcomer but for those who already have infrastructures that they master, it is not the case
- All catalogues are not online and the process to synchronize the download of several datasets can be tricky.
- Platforms should provide more ARD. Essential Variables can be a driver to define which ARD are needed to produce them and make them available. .
- Open data value can be revealed by the use of Web Analytics tools. These data are an opportunity to optimize the catalogues. Unfortunately, the most current free tools are US and their data is not open.

7.2.5 Threats

7.2.5.1 Theory

- Data Security concerns,
- Physical location of hardware is unidentified, therefore Governments consider the storage of their data out of their land and beyond their regulation boundaries.
- Scalability impacts costs that can become opaque in the long run. users need to know when and how long the resources used have been "exceptional"
- Business is highly dependent on the 3rd party Cloud service provider,
- lack of commitment to high quality service and availability guarantees

7.2.5.2 Return of experience from the e-shape pilots

- No problem with security has been reported
- Several pilots had to change platforms and could mitigate the impacts
- It can be necessary to identify where the personal data are physically stored and this information can be difficult to get from the providers
- Web Data analytics are a revealer of the open data value. Currently, the free tools are US and the generated data is not public.

7.3 Synthesis on the usability of Cloud Technologies for Earth Observations - Theory and Practice Status

7.3.1 Lessons learned on usability of Cloud Technologies for all stakeholders

Companies working with EO should develop a strategy for the best use of Cloud technologies for their needs. This strategy will be highly dependent on the size of the company and existing in-house infrastructures. The real cost of the use of existing infrastructures (in particular HPCs) should be considered.

Cloud technologies for Earth Observation require specific training, hiring new staff with these skills, subcontracting experts or getting very good support from providers mastering Cloud and EO.

7.3.2 Lessons learned on usability of Cloud Technologies for EO applications developers

- Developers reaching data download bottlenecks should consider pushing the Application near the Data
- The application architecture should be modular and the component should be containerized in consistent packages in relation to the Cloud resources scalability/elasticity that is needed.
- The operational SLA should be explicit to identify the level of reliability and reactivity of the support. Users should test the reliability (data access and processing) and the reactivity of the support over a reasonable period of time.
- The use of open standards as an enabler to reduce dependency should be encouraged.

7.3.3 Lessons learned on usability of Cloud Technologies for EO Cloud Platforms providers

- Platform providers should keep on offering a free period and sandbox to develop this training on the users' specific needs to identify the technical minimum and maximum requirements.
- Alerts on extra resources activation or threshold of costs and their deactivation should be implemented.
- Dashboards to monitor real resources consumption should be accessible.
- Online catalogues should be over longer period of time, maybe on specific data and coverage to be identified.
- Web Data analytics can be used to optimize the catalogues
- More Analysis Ready Data should be made available
- Essential Variable can be an opportunity
- In situ still driven by the communities. Can ARD on in situ be an opportunity?

8 INFORMATION AND COMMUNICATION TECHNOLOGY- ICT- RESOURCES SPECIFICATION

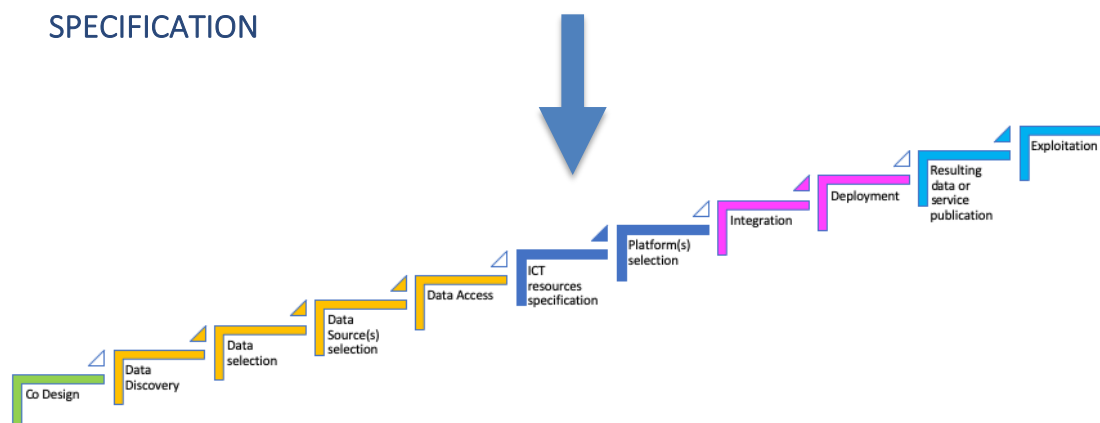


Figure 23: ICT Resources specifications in the Development workflow

8.1 A change of paradigm

In the past, scientists had predefined resources available on their desks or in their computing rooms. They used to adapt their work to the available resources that were available apparently "for free". Maybe they could have liked to work on a bigger geographical area or with a better data resolution, with a better time response but the driver for the trade-off was to fit on the resources that were accessible to them, and they used to constraint their work to fit in.

Cloud technologies change dramatically the paradigm, as the scientist can now define what he wants to do and get access to the resources needed for an affordable cost. So the question is not anymore: How much can I do with my resources, but which resources do I need to achieve my work?

We can observe that when you ask the pilot's developer: which resources do you need?, they get puzzled. Unfortunately, this is the first question to solve to be able to estimate the cost of the resources, to assess the different solution providers relative to your needs, or to anticipate budget requests sometimes very long in advance.

The major benefit of Cloud platforms is their scalability or their elasticity. So to get benefit from them you also need to be able to specify scalable requirements and develop a strategy to minimize the time duration during which you will mobilize the resources. Meaning that this is not one single specification of resources, but several specifications based on different phases of the project or the future use of the system over time.

Moreover, understanding how different cloud resources are used by your application and how they affect your budget, will drive your ICT specifications and underpin the definition of sets of policies you will enforce in different cloud environments. You will very likely Cost Control Policies: identify the resources that impact your budget and release them as soon as they are unused.

These phases of the projects will be defined in order to optimize the Cloud resources needed: first tries with minimum resources, deployment over one cloud to go operational, and eventually change of Cloud provider after some months of operations based on the audience and resources effectively used in operations.

8.2 Sponsorship for Cloud resources

"The Network of Resources (NoR) is an ESA initiative to facilitate the use of cloud environments by users, building on and enlarging the previous Open Science for Earth Observation (OSEO) call, sponsoring R&D users for the use of commercial platform resources.

The NoR call supports research, development, and pre-commercial users to innovate their working practices, moving from a data download paradigm towards a bring the user to the data paradigm, considered essential for maintaining the competitiveness of European data exploitation."

Box 8-1: ICT resources specifications. Lessons learnt from Showcase 4: [MyEcosystem](#) Pilot 1: [mySPACE](#) (better monitoring climate drivers in 25 protected areas).

Given the spatial nature of EO data some level of embarrassingly parallel implementation should be always evaluated. In fact the processing is often done on some spatial neighbour (pixel, object, subtitle) and cloud in High-throughput Processing mode better handles large volume of small job both in term of CPU access and RAM footprint

Recommendations for EO Cloud Platforms providers: Evaluate the possibility to pre-process data to make it easier for applications to subset data. Different options are possible from more low levels (COG tiff, HDF5-like file as Netcdf) to higher level as the different types of data cubes.

Identifying the minimum resources capacities needed for development and testing will facilitate the acceptance for sponsorship

Lessons learned

- The developed algorithms have been implemented in a dockerized environment. This makes the system flexible and easily exportable

References on the Network of Resources:

- Network of Resources: <https://eo4society.esa.int/network-of-resources/>

9 PLATFORM SELECTION

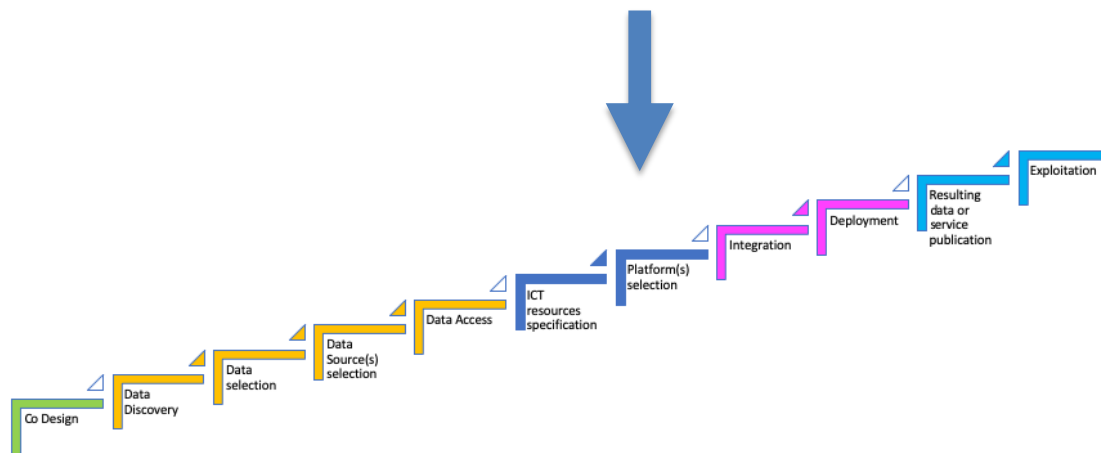


Figure 24: Platform Selection in the Development workflow

9.1 Introduction

In the 80s, Earth Observation developers used to run their applications, models, or processing on personal computers or specialized software and hardware with limited resources and slow response time. In the 90s, faster data transmission rates and more powerful computers allowed the development of distributed EO data processing which enabled collaborative work among researchers accessing and processing data remotely from different locations and sharing resources. Numerical Models were then using all sizes of computers but Supercomputers became for most essential for most of them which required the processing of big data in a limited time as this is the case for Meteorological forecasting numerical models. The investment and running costs of these supercomputers have fostered the development of sophisticated schedulers to optimize the intensive use of technical resources and production constraints.

In the 2000s, the development of cloud computing and virtualization technologies revolutionized EO IT infrastructures. Researchers can now access and process massive amounts of data in real-time using cloud-based processing and storage resources.

Since around 2015, the maturity of the containers technologies, enable the deployment of applications on a public or private cloud platform offering Platform as a service capacity. This has been, in Europe, the start of intensive progress with many initiatives such as the Thematic Exploitation Platforms from ESA, NextGEOSS European project, and the Data and Information Access Services Platforms enabling the Earth Observation community to benefit from the Cloud technologies for data access and processing, to work collaboratively sharing highly specialized libraries, software, and infrastructure, lower the costs when scalability is needed, ...

But these are a lot of evolutions in a short period of time, the implementations are still not fully mature, and before being able to benefit from all the potential of these technological progresses, the EO community needs support.

This is probably why most of the e-shape pilots have expressed, at the start of the project, in 2017, the need to have more information and understanding about the Earth Observation platforms.

Most of the e-shape pilots have expressed, at the start of the project, in 2017, the need to have more information and understanding about the Earth Observation platforms.

9.2 Cloud or High-Performance Computing (HPC)?

Earth Observation applications, Models or transformation processes, can be run on different types of computing resources depending on several criteria but EO, most of the time, relies on very Big Data that often requires Large Scale Computing.

Historically, models and computing demanding applications used to run on "Supercomputers" that, as their name suggests, are super powerful computers. This is what is now called High Performance Computing (HPC). An HPC will typically be capable of quadrillions of calculations (10^{24}) while a personal computer can run some billions (10^{12}).

More recently, more and more companies or consortiums, offer convenient scalable storage and processing capacities in the Cloud at affordable costs, enabling calculations near the data and disrupting the historical model of data download. It enables users to access powerful hardware that they cannot afford otherwise for a limited period of time if needed. The user "rents" scalable resources for a limited time and does not have to care about the hardware or software upgrades or the security of the infrastructure that nowadays requires high expertise skills.

Choosing between HPC and Cloud computing will depend on the volume of data, the required time response, or sometimes the tight scheduling to chain several processes.

For huge amounts of data, HPC computing will outperform cloud computing, as the connection between each "node" of a cloud computing system cannot compete with HPCs. On the other hand, HPCs are far more expensive than Cloud-based solutions but users that really need HPCs can usually access such resources from their IT department or connected Universities.

But even when accessible, the costs should really be assessed as HPC can be far too powerful and more expensive than necessary, when compared to cloud-based solutions. We will discuss later how IT resources budgeting organization probably slows down the adoption of Cloud technologies.

Cloud technology offers to evolve very quickly and these choices should be reassessed regularly to maximize benefits and reduce costs. Cloud technologies mainly promise:

- Flexibility/ Scalability/ Elasticity
- Cost savings
- Security/Disaster recovery
- Mobility/increase cooperation
- Insight / Competitiveness

The Earth Observation Cloud Platform Concept Development Study Report (<https://docs.ogc.org/per/21-023.html>) reveals that satellite data providers are moving towards cloud computing, and implementing the applications-to-the-data paradigm. Right now, the major focus for many providers is to make EO data accessible in the cloud. Others already process, analyse and disseminate EO data in the cloud. Cloud-based systems to store, process, analyse, and make EO data accessible are a paradigm change and disrupt the traditional EO data dissemination and analysis workflow.

Many platforms offer Cloud implementations for the EO community adding EO-focus capacities, libraries, and collaborative spaces, sometimes thematic or not. The Earth Observation Exploitation Platform Common Architecture initiative (EOEPCA) from the European Space Agency (ESA) provides open-source components to connect EO Platforms into a federation. Such a federation, could, in the

future, be an amazing opportunity for the large EO community to access collaborative spaces, EO capacities, and knowledge in addition to the storage and computing capacities provided by the Cloud technologies themselves.

In fact, when budget considerations will be managed correctly, HPC and Cloud Platforms will find their positioning more clearly: HPC will be dedicated to intensive or large-scale Modelling and AI, controlled by the data providers, when Cloud platforms will be used for research, very irregular productions or needs such as trainings, for small modelling or AI triggered by the external users and so requiring scalability and security.

For the user, accessing online pre-processed data or an EO service should not make a big difference. We will see later how it impacts the developments, architecture, capacities, licenses, and business model.

9.3 Selecting a European Platform as a Service Cloud platform

Cloud pioneer infrastructure such as Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure have paved the way towards the democratization of cloud technologies. They have rapidly been instantiated in diverse "as a Service" capacities such as Infrastructure as a Service (IaaS), Data as a Service (DaaS), and Platform as a Service (PaaS) going ultimately towards Anything as a Service (XaaS).

The European Earth Observation Community has upscaled its adoption of these technologies from 2014, with the conjunction of Open Data Policies directives, the availability of Copernicus data and services, and the maturity of the containers making cloud technologies more usable. Building on Copernicus data, several generations of cloud platforms bloomed, enriching rapidly the European Landscape with new capacities and services that needed, indeed, to be characterized to make them more understandable.

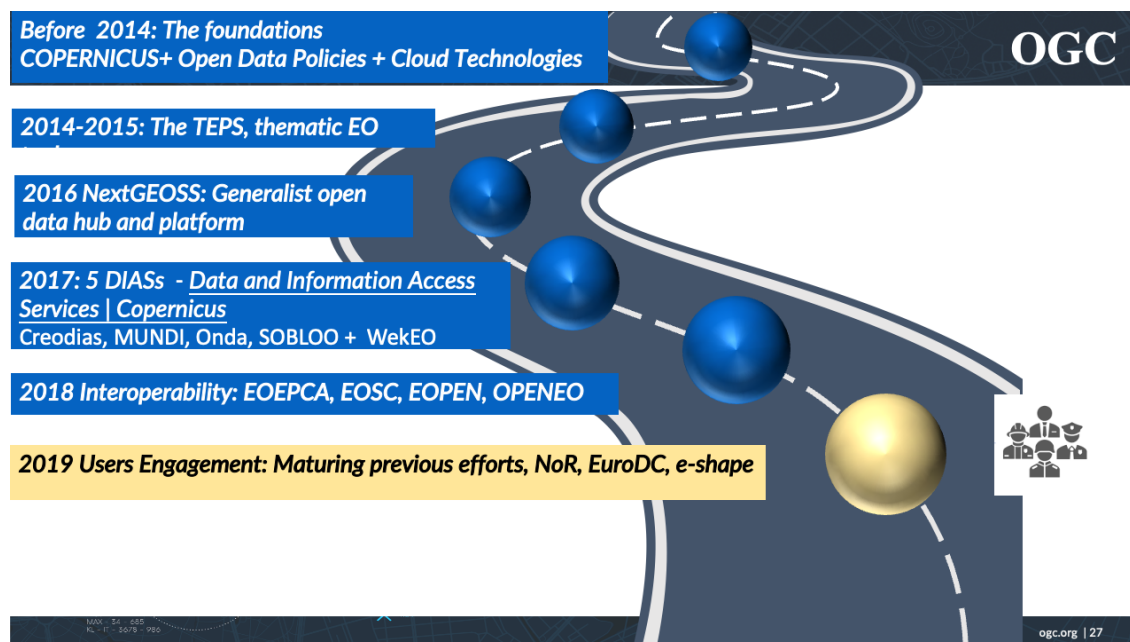


Figure 25: Progress of Earth Observation Technologies in Europe in the last 10 years

Earth Observation is a field where Cloud technologies are particularly relevant in many aspects, and Cloud Technologies can benefit from the Earth Observation community bringing sophisticated scenarios and huge increasing volumes of data.

1. Data storage and management: Earth observation data can generate large volumes of data, which can be difficult to store and manage. Cloud technologies provide a cost-effective solution for storing, managing, sharing and analysing large volumes of data generated by earth observation systems.
2. Data processing and analysis: Cloud technologies enable the processing and analysis of large volumes of earth observation data in near-real-time, providing insights into various environmental parameters such as weather patterns, land use, water resources, and different societal benefit areas such as agriculture, biodiversity, ...
3. Shared collaborative thematic resources: the ESA Thematic Exploitation Platforms, for instance, offer thematic data and related libraries or software to facilitate reusability and collaboration to a community sharing the same thematic interest.
4. Disaster response: Earth observation data can be used to monitor and respond to natural disasters such as floods, hurricanes, and wildfires. Cloud technologies can be used to quickly process and analyze the data, providing timely information to decision-makers to make informed decisions. They can be accessed remotely by different stakeholders to share a Common Operating Picture.
5. Climate change monitoring: Earth observation data can be used to monitor and track changes in the Earth's climate over time. Cloud technologies can be used to store and analyse data, providing insights into the impacts of climate change on the environment.
6. Sustainable development: Earth observation data can be used to monitor and track progress toward sustainable development goals. Cloud technologies can be used to store, manage and analyse data, providing insights into the effectiveness of various policies and programs.

Several European initiatives have therefore focused on the development of Earth Observation Cloud platforms providing Cloud services and technologies tailored to the Earth Observation Community needs, integrating several types of Cloud services such as DaaS, IaaS, SaaS... The experience of the 7 TEPS tends to prove that 5 years minimum are needed for these platforms to reach a good level of maturity. Considering the complexity of the data, the technologies, and the needs, which are all permanently evolving rapidly, this domain is a bit in permanent reconstruction and the diagram below offers a try at formalisation of the status in 2022.

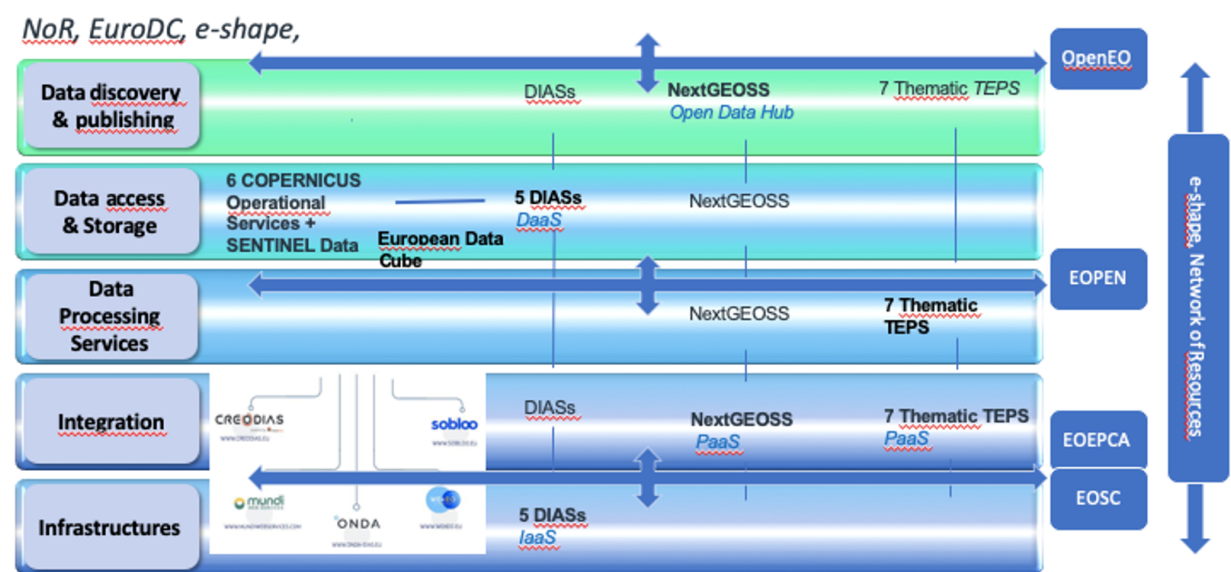


Figure 26: Major European platforms capacities evolution

The diagram shows in bold the initial priority layers of cloud services for the different type of platforms, emphasizing where we can expect them to be the best, and in normal character, the additional cloud layers over which the platforms are currently extending, tending to offer a full stack of services.

More recently, initiatives have focused on the interoperability of these different infrastructures via common architectures or components such as the Earth Observation Exploitation Platform Common Architecture, common interfaces such as EOPEN...

In the context of e-shape, it was important to evaluate each platform based on the specific needs of each pilot and their requirements considering a number of factors, including:

- Data catalogue: Are the data needed as input available or can they be made accessible from the platform?
- Data storage and processing capabilities: Is the platform able to store and process large amounts of EO data efficiently?
- Data and Processing Costs: How much does the data storage cost? How much do the processing capacities cost?

Earth Observation often requires big volumes of data and good performance, So the storage cost policy of the platform can be quickly impacting, all the more that the disk RAID architecture selected to improve the reliability and speed when going operational can contribute to an increased need for efficient storage. E-shape pilots' feedback calls attention to the storage costs that can rapidly become more significant than CPU costs.

- Availability of pre-built tools and libraries: Does the platform provide a variety of pre-built tools and libraries that can be used to process, analyse EO data, or make results accessible? This can help to reduce the time and effort required to develop custom tools and software.
- Integration with other services: is the platform able to integrate with other services, such as geospatial databases, data visualization tools, and machine learning libraries?
- Security and compliance: Does the platform provide robust security measures and comply with relevant regulations and standards, such as GDPR and HIPAA?
- Pricing model: Is the platform cost-effective for the pilot needs and does it provide transparent pricing models that allow estimating the costs of your developments and eventually the production phase accurately?
- Technical support: Does the Platform provider offer support to facilitate the integration, architecture, and cost optimization...?
- Community support: Is there an active and engaged community of developers and users who can provide support and share knowledge?

Such precisions would allow for assessing the risks of using such assets. Clarity is required regarding the basic service provided by the platforms, the APIs, and compliance to open standards. Eventually, the question can be raised on the comparability or unique values of the DIASes and other platforms, to compare their services, performance, interoperability, and ease of access.

Box 9-1: Cloud platform selection. Lessons learnt from [Solar Energy nowcasting and short-term forecasting system \(management support for solar energy plant operators\)](#) Pilot [Renewable Energy Showcase Example](#)

The pilot [Solar Energy nowcasting and short-term forecasting system](#) started operating in a physical HPC environment in an offline mode as a test until the proposed system operating architecture required additional resources including a graphic processing unit and advanced distributed computing. In this direction, the algorithmic architecture was modified in order to be operated in a cloud platform, firstly in WEkEO and then in AWS. In WEkEO the available computer resources were adequate, the system running was seamless and the overall platform reliability for operational EO services was proved as best as possible for cloud solutions.

On the other hand, the cost for a continuation of the service into the long term was unsustainable reaching almost 7-9K euros per month (this cost was not covered by e-shape even for the pilot operating phase). In AWS the cost was reduced to almost 350-550 euros per month but the stability of the platform was not continuous since the availability of the required computer resources was not secured. The final solution in order to reduce the cloud hosting cost was to replace the direct modelling parts with pre-processed ones and to limit the operational capabilities to just the simulation and projection of the pilot data use and production. Concerning the way forward on modern EO services, the optimum scenario is to directly deploy, run and disseminate through the GEO portal including the cloud computing part, the EO data availability, and the configuration to the user community needs in the form of a holistic service creation platform into a competitive cost in order to attract more attention and solution providers.

Box 9-2: Feedback on Using a DIAS Platform for processing of EO data. Lessons learnt from Showcase 5: [Water resources management](#) Pilot 3: [Diver Information on Visibility in Europe](#) (coastal water quality monitoring).

The DIVE pilot selected CREODIAS as the DIAS provider and has made the following observations/lessons learned:

- The platform as a means of hosting virtual infrastructure had good stability with no noted downtime
- Charges for using the service were reasonable for what it offered.
- Adding credit to the service was a time-consuming process unless you had a credit card ready to pay. We had to pay invoices but there were delays in getting payment confirmed.
- The support was good in terms of responsiveness when raising new tickets for problems that occurred during our use of CREODIAS
 - For example, It was easy and quick to raise new tickets and they were acknowledged quickly.
 - Support was also not good because although the tickets were raised quickly and were responded to, the resolutions to these problems took a long time (e.g. up to 2months in one instance)
- EO Data availability was inconsistent and unreliable. We had several incidents where our selected datasets were not available for sometimes a couple of months at a time. The last incident was because the data provider had made a change but CREODIAS seemed unaware of the problem which means a change would be needed on their side and this was never implemented to resolve the problem.
 - This would have been a bad solution to choose, had this been in a production environment or where a service had users which were paying for the service.
- EO data availability was unreliable to the point that a backup solution to pull data directly from the provider via an API
 - Our pre-processing we were running on the DIAS was eventually migrated to be hosted in-house as CREODIAS was no longer able to provide our dataset and we had lost confidence in the service.

Lessons learned

CREODIAS has a good selection of EO data available and our pilot didn't need any extra requirement to download any additional data to do any processing, it was too unreliable (in terms of data availability) to utilize for anything other than development/research purposes.

We would not recommend this particular provider for anything other than development/research purposes, at least based on the problems we had, as we now run our processing locally and pull data from the data provider's API. This final solution provides several benefits as our pre-processing is successful and we do not incur costs charged for using cloud provider resources.

For any solution selected, we encourage users to clearly express their operational expectations and review the Service level offered by the platforms involved in the architecture.

10 INTEGRATION

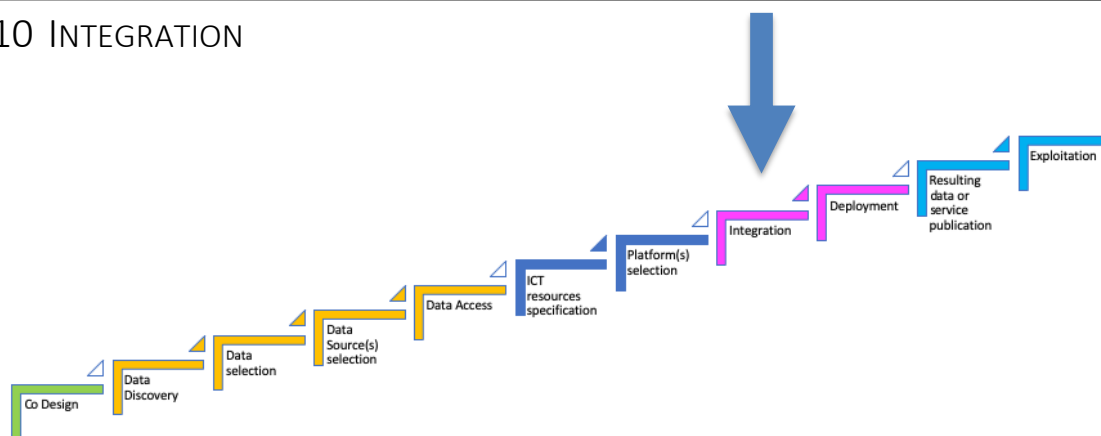


Figure 27: Integration in the Development workflow

10.1 Open source tools

Open source tools have been increasingly used in the field of Earth observations in recent years, providing researchers, scientists, and other stakeholders with access to powerful, flexible, and customizable software built on high-level expertise and skills related to observation and measurement processes, data preparation, data processing, data analysis, data visualization and software deployment. The catalogue of open-source tools used by the e-shape pilots is provided in Annex 5.

The use of open-source tools in Earth observations has several benefits, including increased transparency and collaboration, reduced costs, and increased access to powerful tools for processing and analysing data. Open-source tools also allow for customization and adaptation to specific research or application needs, making them highly flexible and adaptable to different contexts. In the context of GEO it is also important to enable open reproducible research.

10.2 Containers

Integration requires being aware of the level of dependence and risks of becoming platform dependent, especially when using capacities provided by the platform to enable elasticity.

Cloud containers are a technology that allows software applications to run in a consistent and isolated environment, regardless of the underlying infrastructure. Containers are lightweight, portable, and can be quickly deployed and scaled up or down as needed, making them a popular choice for cloud-based applications.

The expected benefits from cloud container technologies are:

- **Portability:** Containers are designed to be portable and can run on any infrastructure that supports containerization, such as public or private cloud platforms. This enables moving environments, such as from development to integration, from integration to validation, and from validation to production, or from one cloud provider to another. Several e-shape pilots have changed platforms for different reasons during the project.
- **Scalability:** Containers can be easily scaled up or down depending on demand, allowing for more efficient use of resources and better performance during peak periods. This also means that if several components of the application have different scalability needs, such as a front office on the web that will have to scale as the audience increases and the back office that will have to scale if the input data volumes increase, this different component will have to be packaged in different containers.

- **Resource efficiency:** Containers use fewer resources than traditional virtual machines because they share the host operating system kernel. This makes them faster to start up and shut down, and more lightweight than virtual machines. Nevertheless, containers are designed to run on specific operating systems, and some operating systems are not well-supported or fully compatible with containerization technology. For example, applications that require a specific version of the Linux kernel may not be able to run in a containerized environment. Containers also still require resources to run, such as CPU, memory, and storage, so may not be the most efficient solution for applications requiring significant amounts of resources.
- **Consistency:** Containers provide a consistent environment for applications, ensuring that they run the same way across different environments and infrastructures but they can add complexity to the development and deployment process. As mentioned in Annex 5, 10 e-shape pilots have mentioned using Docker, meaning that the technology is quite well adopted but there can be challenges associated with container orchestration and management or network complexity, particularly for larger organizations with complex IT environments. In this case, deployment can require involving Network experts.
- **Security:** Containers provide a level of isolation between applications, reducing the risk of security breaches or conflicts between applications. But this requires to be properly configured or managed to avoid for example risks of cross-contamination and data breaches if containers are not properly isolated from one another. Additionally, vulnerabilities in container images can expose organizations to security risks.
- **Flexibility:** Containers can be easily customized and adapted to specific application needs, allowing for greater flexibility and agility in development and deployment.

Overall, despite some potential disadvantages to consider in the areas of complexity, network architecture, resource constraints, security, and compatibility, and even certain legacy applications or applications that require specific hardware configurations cannot be containerized, containers offer several advantages that are particularly important and have been useful to many e-shape pilots.

Box 10-1: Using Dockers to develop an API for DIVE. Lessons learnt from Showcase 5: [Water resources management](#) Pilot 3: [Diver Information on Visibility in Europe](#) (coastal water quality monitoring)

The DIVE application offloads most of its processing and data management to a server that is accessed through a public API. This minimises the load on the mobile application itself and allows us to use the tools of our choice to do the heavy lifting in the API.

PML has built a large number of web services in a variety of languages generally hosted on PML's own infrastructure. In this case we were keen to explore other development options that might allow us to scale easier or deploy the API service on other platforms.

Docker

We decided to try Docker (<https://docs.docker.com>) for a number of reasons.

- The container-based system would be easy to deploy on our VM infrastructure and to move to an external host if we wished.
- By using prebuild docker images for support services such as the database, we could avoid any additional software requirements on the host server.

To make life easier we used docker-compose (<https://docs.docker.com/compose/>) to manage the various docker images that made up the system. This meant we did not have to worry about individual container names and ids as docker-compose would start up all the required software as defined in its configuration file.

Development process

The development process itself was largely unaffected by using Docker. The only negative impact was the introduction of a docker-compose down, build, up cycle every time a change was made. This could have been

avoided by using additional tooling such as some of the plugins available for Visual Studio Code (<https://code.visualstudio.com/>) but the rebuild time was so short that this was considered unnecessary. On the positive side the containerisation made moving between different development machines and the live server very easy.

Deployment

As we were developing and running locally, we did not make full use of Docker as a deployment tool, the live system was built from source on the server itself. This probably was not the best way to do it and we have now produced a deployable Docker image with supporting files that can be run without the build stage. This is publicly available at <https://github.com/pmlrsg/diveapi>.

Conclusions

- Docker was easy to install and deploy (on Linux).
- Our development process was largely unchanged by using Docker.
- Containerisation made the service easier to deploy on our VM infrastructure.

We would use Docker on similar projects in the future.

11 DEPLOYMENT

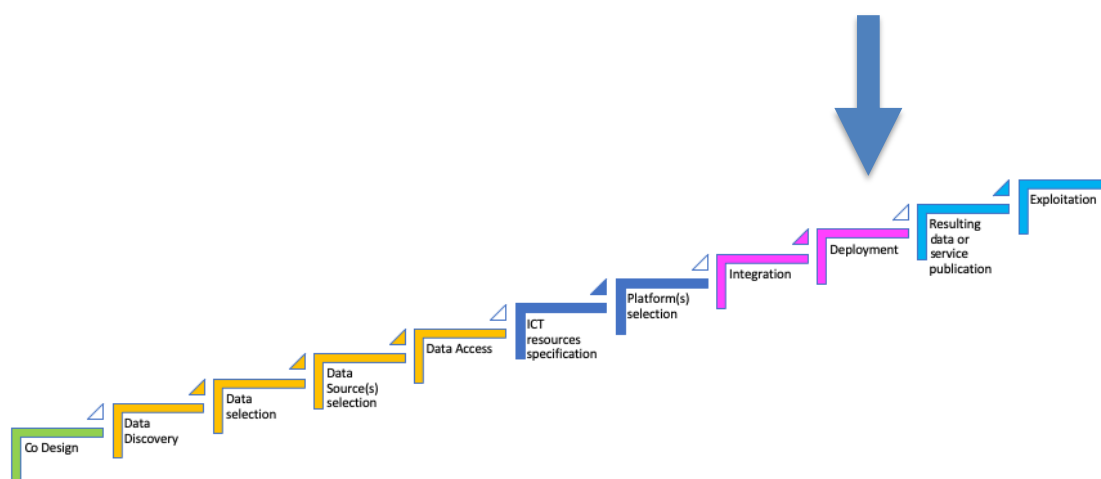


Figure 28: Deployment in the Development workflow

Once the application is developed, tested, and debugged, it has to be deployed. This requires a review of each of the components and deploying an Earth Observation (EO) application will typically involve:

- Infrastructure setup: Deploying an EO application requires setting up the necessary hardware, such as servers but also electrical supply, software such as databases, and networking components.
- Data access and management: This involves acquiring, storing, and managing the EO data that the application will use, such as satellite imagery, in-situ data, and environmental or weather data or deploying the application near the target operational source of data,
- Data preparation and processing: Most of the time, the input data is not in the shape needed by the application, and some data preparation is required including accessing, processing, and analysing the EO data to extract the necessary information that will eventually have to be stored to be made for the application. Efficient Analysis Ready Data standardization could lighten up this phase,
- Integration and testing involve integrating the various components of the EO application, testing its functionality and performance, and addressing any issues or bugs that arise during testing. This can

be a critical step for applications on the web or mobile applications that are expected to have a huge audience. Rigorous load tests with specific tools such as JMeter, have to be made before opening the application to the public.

- Deploying the EO application to the production environment has to be planned and coordinated with all the relevant teams including points of contact from the infrastructure, software, and data providers but also network experts to make any necessary updates or fixes to ensure that it continues to function properly. It will be much more challenging if the application is already operational and has millions or thousands of users or if it is critical for security obliging to make it without any interruption.
- Monitoring its performance and usage has to be organized carefully: monitor the disk spaces (logs can grow very fast), monitor the RAM and CPU use, set alerts on critical thresholds for all these monitoring, check the presence of critical background applications, check the accessibility of each external components at any time, ... Web analytics tools can contribute to this monitoring or to anomalies analysis.
- User training and support: This involves providing training and support to end-users of the EO application, such as scientists, researchers, or government agencies, to ensure that they can effectively use and interpret the results of the application.
- Exploiting team training, support, and operational protocols, guidance, or documentation including an operational "cookbook" with all procedures to be activated in case of problems.

Deploying an EO application can be a shorter step in terms of duration than the development but it is critical. It can involve a large number of actors with different skills that will have to be booked and coordinated to be able to solve the problems rapidly to ensure that the application meets the needs of its users and delivers accurate and reliable results on time.

When dealing with critical applications, it may be necessary to use anti-fragile strategies or implement multiple redundancies or platforms to ensure their reliability. In such situations, a decision must be made about whether to deploy all the application components on one platform or distribute them across multiple platforms. It can be more reliable and less expensive to deploy on different platforms implementing different solutions with a load balancer than asking for a very demanding Service Level with an extremely high operating rate.

12 RESULTING IN DATA OR SERVICE PUBLICATION OR DISSEMINATION

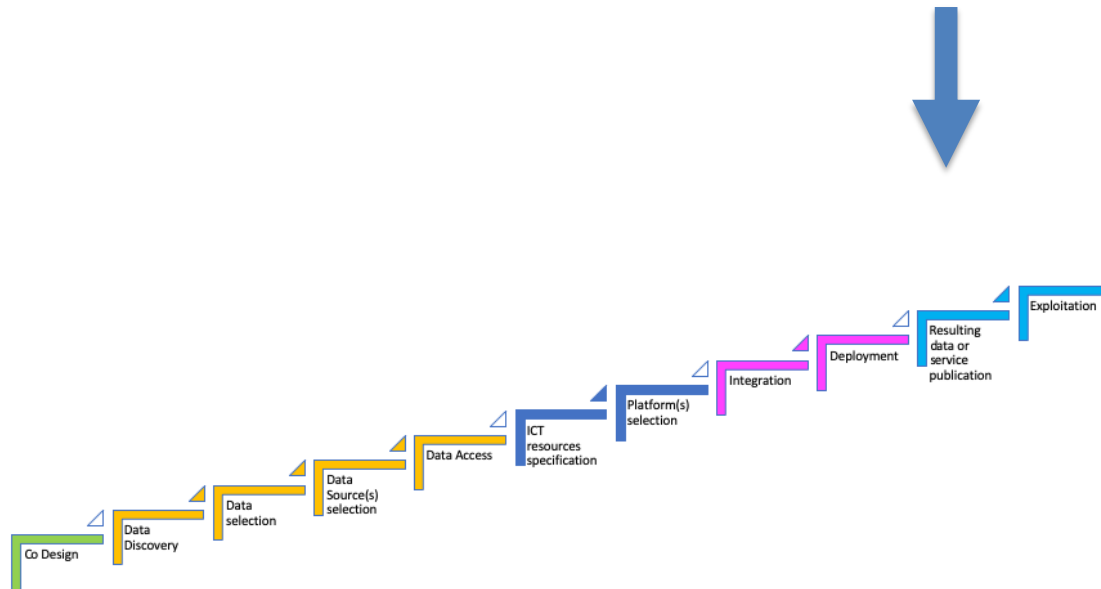


Figure 29: Resulting Data or service publication in the Development workflow

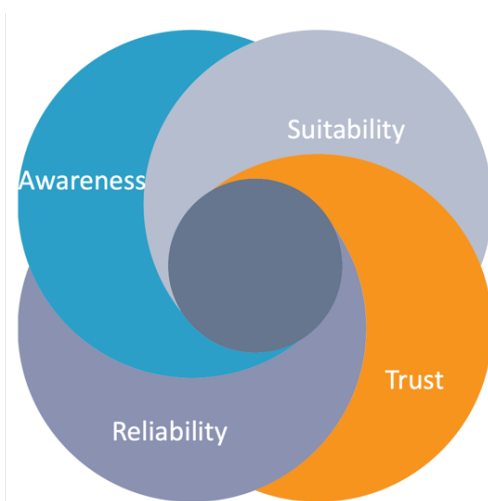
12.1 Introduction

As introduced in the 'Data, Process, and Application Discovery' chapter, the eminent element of the EO user's journey is finding the best matching resources. That is why the supply of the products shall close with the discoverability of outcomes.

As cataloguing gained interest already several years ago with the start of equipotential EO data growth, a number of approaches have been proposed to reach out to various groups.

Considering the value in the richness of choice, it is important to use available tools for proper aims. Therefore, herein the chapter outlines the publication options and best practices to guide through various aspects of the process.

12.1.1 Impact through presence



Maximizing impact requires targeted dissemination to appropriate audiences, in the appropriate context, accessible (e.g. process and format), with the right messages, and with the right license.

Both inbound and outbound marketing strategies shall be supported by content that is truly findable and available on the web.

Recognition always starts with awareness building and convincing on the suitability of products. This is the moment general information is published or actively distributed among our potential target groups.

Here, the key factor is the informativeness of the message and its position on the user surfaces (like search engines, databases, and knowledge bases).

Persistent content on the web helps with both positioning and word-of-mouth dissemination. It also supports the initial trust credit one can be granted.

But the coherence between promise and delivery is what strengthens the relationship built on reliability or confusion and triggers looking in other directions for help.

Considering the specific EO positioning in the target markets, there is still space for activities based on the dissemination of value-building uptake and sustainability of the business:

- building awareness of capacities and capabilities of EO,
- increasing capacity contributing to the general knowledge including open access to various publications, related sample data, and tools,
- discovering insights from the uptake like traffic measurements, feedback, requests, and complaints that help to identify needs, explore the ecosystem with its potential, limitations, and values,
- building trust and reliability through coherent messages on multiple levels.

Selection and implementation of the above aims shall be integrated with the overall marketing strategy. For example, services and datasets catalogued have an increased chance for uptake already. Platforms like GEOSS, NextGEOSS, and Google Datasets are continuously visible on the thematic panels and these aggregators allow for wider outreach within various target groups.

However, just being listed is not the optimal leverage of uptake like having the webpage indexed by the search engine is not guarantee the audience will flock to one.

A proposed efficient way to implement persistent on the web shall start from the following questions:

- what platforms shall be considered based on the target and information relevance,
- how much data should and can be shared,
- how to publish content and possibly,
 - how to integrate,
 - how much effort is required and how to minimize this effort.

12.2 Publication platforms overview

Data dissemination platforms considered in this chapter can be separated into three categories, each having its characteristics:

- data and services access brokers - NextGEOSS, GEOSS, Google Datasets
 - Focus: (meta) Data accessibility is their main objective
 - Target: mainly researchers
 - Input: metadata describing resources is required
 - Added value: facilitate INSPIRE, FAIR, and GEO Data Management Principles
- exploitation and research platforms - Thematical Exploitation Platforms (TEP-s), Data and Information Access Services (DIAS-es)
 - Focus: exploitation of EO data, third-party data, and services are their added value
 - Target: researchers and business
 - Input: organization and use case information, data, and services

- targeted business outreach
 - Focus: play EO business Yellow pages and markets role
 - Target: EO downstream market
 - Input: organisation, use cases, success stories information

12.3 Data and service metadata

While focusing on the data and services dissemination (points 1 and partly 2 from above) the best practice vary based on the character of the data or services that one is going to publish.

e-shape Data Management Plan provides DMP and FAIR principles exhaustive considerations and it is recommended, to start the publication process, to analyse how the metadata is going to be provided.

Metadata is based on international standards.

Metadata can be provided as:

- static file describing data resource, catalogue, or service. It can be provided as input for dissemination platforms or exposed online. As static files, they cannot contain detailed information about specific data resources. Best for homogenous data sets where characteristics (metadata) do not change often. Metadata can be generated using available tools (see NextGEOSS). Examples:
 - sample data resource metadata
 - sample catalogue metadata
 - sample service metadata
- metadata catalogue exposed by the web application. Best for data sets with resources of various characteristics (for example various products, coverages, time). Online metadata is available based on the data resources available in the repository. Open-source example implementations:
 - GeoNetwork - <https://geonetwork-opensource.org>
 - PyCSW - <https://ckan.org/>
 - CKAN - <https://ckan.org/>

Note: the guideline does not consider data hosting solutions, while some hosting services (i.e. ARMINES) provide already a data check-in process that fulfils mentioned principles and ensures the hosted data is available on the brokering and dissemination platforms.

12.4 Defining a data license

12.4.1 Introduction

A license provides clarity and certainty on possible downstream usage of Earth Observation services, which enables innovation for research, and business and supports its sustainability. A license depends on the business model, but it also depends on the licenses attached to the input data. Crediting the data used is requested for instance by Copernicus and can be critical for some data providers.

Being compliant in the licenses management will encourage data users to share their data because credit is mentioned as critical for many data providers. There is a global consensus that we need more global in-situ networks and data collections and the European Commission as well as GEO are currently advocating for more in-situ data sharing. In-situ comes from a very fragmented community to whom it would be easier to advocate data-sharing if we could recommend some best practices and if the licenses were then correctly implemented.

Another topic that makes licenses more critical than before is the use of Artificial Intelligence. AI can enable the merge, fusion, analysis of huge amounts of data, but the more data used, the more licenses have to be mastered correctly.

To sum up, it is important to provide clear and as simple as possible licenses best practices and to implement them rigorously to foster data sharing particularly for in-situ, to increase data uses including for business, to support the transition from research to operations, to empower Artificial Intelligence, Machine learning and other new technologies.

The diagram below aims to address in a non-exhaustive way, the importance of licenses attached to output data and how these can impact the business of a product.

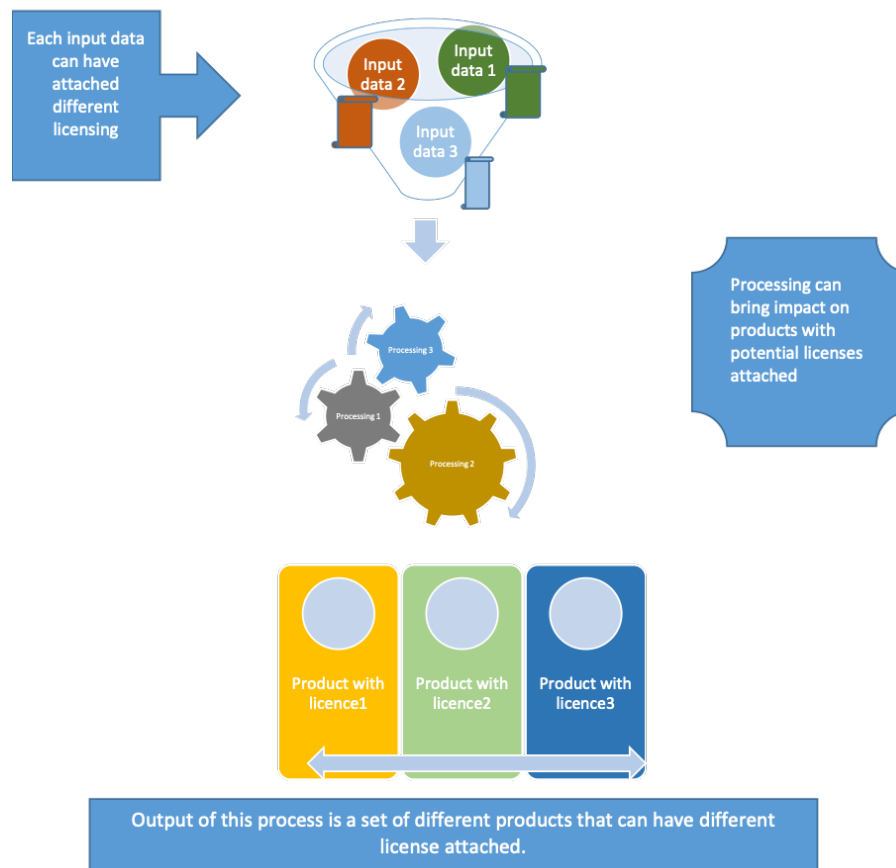


Figure 30:Inputs and outputs Licenses

The outcome of this process is the ability to attach different data licenses to different products, which can then be displayed on various portals to showcase open and free data for multiple purposes, such as research and business.

Within the e-shape project, the majority of pilots have produced datasets, services, mobile applications, open-source code on GitHub, and documents, which have been released to the public. To increase the usage and users of EO services, it is important to clearly attach a license to these outcomes.

This section on licensing aims to provide the EO community with a better understanding of the importance of licensing for the uptake of EO products in the market.

The results of the e-shape pilots have been published, disseminated, and potentially commercialized. To promote their solutions and outcomes, pilots may publish them on different portals targeting multiple communities and increasing their visibility or making them discoverable via Search Engines.

However, legal aspects, such as licensing, cannot be delayed and play a crucial role in this final pilot phase.

Licensing is one of the many legal aspects that must be considered for the commercialization of pilot results. Attribution and giving proper credit to the source is commonly requested for images on the web, and the same is expected for the Copernicus programme, which requires proper credit when Copernicus data is used for the processing of EO products. This is just one example of the many complex topics that e-shape pilots have been aware of in the world of licensing.

Below, we summarize some of the main aspects of data licensing and provide best practices drawn from this work to benefit the entire EO community. Services Licensing approach can be different as the licensing is not addressed in the same way for data and for applications. Efforts have to be made to simplify these licensing aspects to enable the usage upscale. It should be the same for the user if he uses pre-processed data or a service processing the data on the fly.

12.4.2 Copernicus' open data policy

The European Commission conceived Copernicus' data as full, free, and open to allow the scientific community and developers to use Sentinel data and other Copernicus data without legal restrictions. The goals are to enable science to take full advantage of the value of Copernicus and to foster the development of businesses. By "no legal restrictions," we mean that users can obtain Sentinel and Copernicus data without paying any fees and can distribute, reproduce, or publish from the source or data provider, which is the European Commission.

12.4.2.1 Conditions

When using Copernicus data such as Sentinel data to create products, developers or individuals must indicate that the product has used Copernicus data. Regardless of whether other sources besides Copernicus data were used, the attribution legend must be stated in the product. If the developer modifies the data and creates value-adding information, they must still reference Copernicus as the source data in the modified data.

12.4.2.2 Restrictions

Although Copernicus' open data policy allows as much freedom as possible to use its data, the European Commission imposes restrictions, especially regarding data that could impact the security of the Union's Member States. Therefore, access to certain types of Copernicus data, mostly High-Resolution data, is restricted to predefined users by the EC.

12.4.2.3 Warranty of Copernicus data

The terms and conditions of Copernicus' open data policy do not include a warranty clause, which means that the European Commission cannot be held liable for any damage caused by faulty input data or information.

12.4.3 General aspects of EO data licensing

12.4.3.1 Ownership of EO data or copyright

When creating an EO data product, ownership rights can vary depending on the terms and conditions outlined in the product's license agreement. It is crucial for developers to determine if their data is protected by law or if they need to safeguard their ownership rights in the data that will be available to the end user. Typically, developers achieve this protection through contractual arrangements, such as licensing agreements for the EO product. It is worth noting that developers aim to safeguard the value-added product resulting from their processing, which includes intellectual creation, rather than the input data collected from satellite sources. This protection can extend to all the contents of a database,

which may be subject to sui generis rights. Under these rights, the database maker is entitled “to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database”.

12.4.3.2 Usage rights

In order to grant the end user rights to use the developer's data or product, it is necessary for the developer to specify which types of usage are permitted. This includes whether the user can simply view, download, save, copy, reproduce, process, modify, merge, integrate the data into internal products and applications, integrate it into business processing for third parties, integrate it into their own services for third parties, or any combination of these options.

Another important consideration is the purpose for which usage rights are granted or prohibited, and whether the right to use the data is perpetual or temporary. Some pilots have stipulated that data will be open for the duration of the e-shape project. Clear drafting of usage rights and obligations is crucial for providing legal clarity to both the user and the developer, particularly in the context of developing datasets, services, or mobile applications for the pilots.

12.4.3.3 Warranty and Quality

To ensure clear expectations and responsibilities between the licensor and the developer, it is important to specify these in the licensing terms and conditions. The level of responsibility of the developer can depend on the quality and accuracy of the product offered, as this can impact the expectations of the end user. This, in turn, can impact the drafting of any warranties or disclaimers.

However, there is currently no standardization in licensing terms when developing datasets, services, or mobile applications for the e-shape pilot. This means that accessing various EO data sources may involve a range of licenses, particularly if the input data is from private sources, which can impact the pilot's license. As a result, legal interoperability of the licensing terms can pose significant challenges for the pilots.

12.4.4 GEO Data Licensing Guidance

In February 2023, the Law and Policy Subgroup of the GEO Data Working Group recommended the use of the e following data licenses, consistent with the GEO Data Sharing Principles: •

- Creative Commons Zero 1.0 Universal Public Domain Dedication (CC0)
(<https://creativecommons.org/publicdomain/zero/1.0/>) •
- Open Data Commons Public Domain Dedication and License (PDDL) v1.0
(<https://opendatacommons.org/licenses/pddl/1-0/>) •
- Creative Commons Attribution 4.0 International (CC BY 4.0)
(<https://creativecommons.org/licenses/by/4.0/>)

GEO Members, Participating Organizations, and other entities that share open, unrestricted data should clearly license such data using only one of these licenses. Custom license agreements should not be used, and these standard licenses should not be modified or augmented with additional text.

12.4.5 Conclusion

Licenses should be viewed as a valuable tool for developing a business model and not just a burden. By attaching a suitable license to their products or services, developers can effectively scale up their usage and reach a wider audience. In the context of the e-shape pilots, licenses can provide legal certainty to their activities and ensure a smooth commercialization process. A well-crafted license for EO-based

products or services can provide clarity and certainty not only to the data owner but also to the end-user, making it an important consideration for businesses while developing their models. Licenses also play a crucial role in the FAIR-4 principle of increasing data re-use by clarifying licenses, as demonstrated by the DMP self-assessment tool.

References:

- Copernicus Copyright and licenses: <https://www.copernicus.eu/en/access-data/copyright-and-licences>
- GEO licensing guidance removes barriers to open data sharing https://www.earthobservations.org/geo_blog_obs.php?id=590
- The Legal Side of Open Source: <https://opensource.guide/legal/>

12.5 Publishing on the web via data portals

12.5.1 Introduction to publication via major data portals

The e-shape project has advocated and supported the publication via several portals to reach a bigger audience, eventually new markets revealing the value of the pilot's outcomes and of the European resources they were built on. Several pilots emphasized that they could not spend time on this diversity of publications and it should be as easy as possible to optimize the workload. This is why the project has invested time in facilitating the publication to the GEOSS DAB and the GEO Knowledge Hub, asking for all the needed metadata and going through the Web Service Energy (<http://www.webservice-energy.org/>). All the e-shape pilots are discoverable via this portal that is harvested regularly by the GEOSS DAB and the paragraph 12.5.3 presents how this work has enable the publication of Knowledge Granules in the GEO Knowledge Hub.

In order to go beyond these major portals and give more autonomy to the pilots to extend their publication in NextGEOSS, the DIASs and other community portals, information has been collected about the standards and/or processes useful to further publications.

12.5.2 Publication platforms details

Due to the dynamic and rich ecosystem, the below description tries to present as much information as possible, but does not aim to be exhaustive in a horizontal view.

Table 4: Publication platforms details

Platform	Information type (what shall be visible)	Process (how to be visible)	Externally indexable (what engines index and link to the web resources)	information set	target user	use case	interfaces	limitations
NextGEOSS	service	<ul style="list-style-type: none"> register on platform fill the form with metadata send through ServiceDesk 	general search engines	ISO19115 based	EO business	Service advertisement	ISO19139 file upload	
					EO specialists/researchers	Service endpoints exposure		
	data from the collection(s)	<ul style="list-style-type: none"> expose harvestable interface configure harvester 	Google Datasets	ISO19115 based	EO specialists/researchers	Standardised data uptake increase through NextGEOSS and GEOSS	OpenSearch on Atom/JSON custom (through CKAN extension)	
	data/data collection	1. register on platform	Google Datasets	ISO19115 based	EO specialists/researchers	data endpoint exposure and linking	any web URL	data resources can be visible as



Platform	Information type (what shall be visible)	Process (how to be visible)	Externally indexable (what engines index and link to the web resources)	information set	target user	use case	interfaces	limitations
		2. fill the form with metadata 3. support Service Desk in Registration						catalog, without access to specific records.
	data from collections	1. request integration 2. implement and test the CKAN extension	Google Datasets	preferably ISO19115	EO specialists/researchers	Increase data uptake of custom (non-opensearch) catalog	Any	The integration is costly and not preferred without additional resource allocation.
GEOSS	data from the dataset OR whole collection, service	1. expose interface 2. register - fill out the form[link] 3. coordinate with Yellow pages process	NextGEOSS (though specialised search)	initially <u>simplified survey</u> ,	EO specialists/researchers	Increase data update, certify quality as GEO labeled data provider	over 50 interfaces including OpenSearch, CSW, FTP, CKAN	



Platform	Information type (what shall be visible)	Process (how to be visible)	Externally indexable (what engines index and link to the web resources)	information set	target user	use case	interfaces	limitations
Google Datasets	collections	entail the pages and register on google for crawling	Google search	schema.org/W3C DCAT/W3C CSVW	EO specialists/researchers	Increase organisation and datasets visibility, increase uptake.	crawable web pages	
EOMall	entity, service	1. ask for publication 2. describe your service and provide high-quality graphics	general search engines	org name, type, www URL, activity class and type, keywords, description (multipart), logo, images	EO (related) business			EOMall is not providing graphics, that

12.5.3 Publishing on the GEOSS Portal

12.5.3.1 GEOSS features basics

GEOSS (Group on Earth Observation System of Systems) is one of the GEO foundational tasks aimed as an access point to all the community resources. While GEO Resources registered on the Portal are then available through the web portal search and discovery functions and programmatically via GEO DAB API.

The GEOSS infrastructure supports both periodical metadata harvesting and distributed queries. The web portal is equipped with visualisation and dynamic filtering functions so the user is able to preview the footprints but also the data on the map as well as use transformation functions to download freely available (see Manual).

The resources with OGC WMS endpoints can be overlayed directly on the discovery interface. Similarly, data from OGC CSW can be downloaded in several formats using the portal functions.

Each Data Provider is listed on the portal Yellow pages and by the resources.

All the data resources are welcome on the platform, while GEO Data Management Principles compliance is encouraged.

e-shape is a contribution to EuroGEO which is a European contribution to GEOSS. Advocating and engaging with GEOSS is at the core of the project. As such making data and applications discoverable via the GEOSS portal and accessible via the GEOSS portal or the GEOSS API is a target for all pilots.

The registration process supports a variety of interfaces including simple FTP (see Manual), but interfaces providing normalised metadata (like CSW, OAI-PMH etc.) make faster integration.

GEOSS offers also the so-called Reuse Components to serve the specificities of the various user communities. The Reuse Components are:

- The GEOSS View, which provides access to a subset of specifically defined GEOSS resources using temporal, thematic, and spatial criteria;
- The GEOSS APIs, which expose the discovery and access functionalities of the GEO DAB and as such can be exploited by user communities' client applications or portals;
- The GEOSS Mirror is a GEOSS Portal site customisation for SBAs, Flagships, Initiatives, and Communities. The customisation better serves the specific community interests by filtering catalogues and search results by a specific theme or GEO DAB view, location of interest, etc.;
- The GEOSS Widget is a freely-available instantiation of selected GEOSS Portal widgets made available for possible customization in various areas of application (e.g. a specific SBA, Initiatives, etc.). This is accomplished by publishing portal code parts (widgets) wrapped up an API." (<https://www.geoportal.org/community/guest/general-information>)

12.5.3.2 GEOSS publication process

Registration starts with an initially simplified survey where the basic information is collected including resource URL, description and accessibility, provider and contact data, and declaration about DMP, SDG/SBA relevance.

Once registered, the data provider is contacted and guided by the GEOSS Maintenance Team which integrates data in the catalogue.

Within the process, DMP compliance is evaluated using a number of dedicated tools, and the Data Provider is advised on how they can be leveraged.

The process looks at the metadata quality and completeness, data accessibility, and reliability of the service endpoints.

After the evaluation and acceptance, the resource or repository is considered compliant with DMP.

12.5.3.3 Further information

For comprehensive information about the process see:

- GEOSS up-to-date information package - <https://www.geoportal.org/community/guest/documentation>
- [Manual n. 1 All you need to know to become a GEO Data Provider](#)
- [GEO DAB \(Discovery and Access Broker\): Registration Guidelines](#)

Box 12-1: EO-Based Surveillance of Mercury Pollution Services and data resources discovery and access via the GEOSS DAB catalogue after direct publication of the Pilot's outcomes

[EO-based surveillance of Mercury pollution](#) (Minamata Convention) Pilot - [Health Surveillance](#) showcase has published its resources as 3 services and 1 data resource.



Box 12-2 High photovoltaic penetration at urban scale pilot's metadata details in the GEOSS DAB after harvesting of the metadata from the GeoNetwork implementation from the WebService Energy

The S3P2 pilot resources are made available as GEOSS browses the instance from GeoNetwork and harvest the metadata

Resource detailsShow raw metadata

e-shape Pilot S3-P2- High photovoltaic penetration at urban scale

Rooftop PV systems in urban areas are very interesting because they do not emit air pollutants nor GHGs during their exploitation, they produce electricity where this electricity is consumed and they add value to unused urban roofs and reduce urban heat island effect. But, due to complex shading effects in urban context (vegetation, surrounding buildings, superstructures of roofs, etc.) and local atmospheric and meteorological effects, their massive penetration in urban areas will induce a significant variability in space and in time in the energy injected in the electric grid. As far as the electric demand side is concerned, a detailed modelling of energy requirements from residential, commercial and industrial buildings with varying demand profiles for electricity is also required. Therefore there is a need, in urban area, for GIS-like tool for grid operators, urban planning decision makers, industries, aggregators for solar energy trading, citizen (PV self-consumption) and researchers. This GIS-tool is meant to provide an urban energy system modelling of distribution grids to plan, monitor and nowcast (i.e. and short term forecast) the spatiotemporal variability of the electric consumption on one hand and of the production of fleet of PV rooftop systems on the other hand.

Contact information

Contributor: **Lionel MENARD**
Delivery point: -
City: **SOPHIA ANTIPOLIS**
Postal code: **6904**
Country: **6904**
E-mail address: -
Organization name: **MINES ParisTech / ARMINES**
Role: **SOPHIA ANTIPOLIS**

Data identification

File identifier: **d9f3542d-2626-4294-b61e-c356faf0741f**
Parent identifier: -
Hierarchy level: -
Date stamp: **2021-03-05T17:40:07**
Language: -

12.5.4 Publishing on the GEO Knowledge Hub (GKH)

In the framework of e-shape DoW, the impact number five (I-5) and the corresponding assessment indicators of the e-shape project aimed at addressing a "*Coherent data management, through the use of GEOSS Data Management Principles and best practices (INSPIRE-compliant)*". A sample of such practices has been achieved within e-shape, through the publication via the webservice energy catalogue, harvested by the GKH and the GEO DAB. They were mainly focused on ISO 19139 Metadata records and their compliance with the INSPIRE directive. Metadata records exemplifying such achievements are available on the web service-energy GEO Community Catalogue: <https://tinyurl.com/e-shape-Pilots>

These metadata records have been built in the spirit of the knowledge granule concept as supported by the GEO Knowledge Hub (GKH) platform: <https://gkhub.earthobservations.org/>

All e-shape Pilots are available through the GEO KH, harvested from the webservice-energy catalog at: <https://tinyurl.com/GEO-Knowledge-Hub>

Looking in detail at the general structure and the available "Elements" of the Knowledge Package", it can be seen that it imports the elements available on the original ISO 19139 metadata record created on the web service-energy catalogue.

The Knowledge Package, as offered in the GKH, is providing extra possibilities including Digital Object Identifier (DOI) attached to the granule resource, SGD (Sustainable Development Goals) classification support, Target audience and GEO engagement priorities definitions, Versioning, and if needed upload repository support.

Automatic harvesting from the GHK onto the web service-energy catalogue was developed (Powered by the open-source GeoNetwork tool) to automatically render a complete GKH package from the original ISO metadata record. This approach mimics the current efficient and well-known workflow of the GEO DAB (Discovery and Access Broker) adding the extra possibility for GKH data providers to re-edit and manage the harvest process of their own resources in the GKH.

One of the main lessons learned along e-shape is the opportunity to deliver scientific data to the public. Few Pilots have experienced the workflow from data sharing to knowledge sharing on their own.

Box 12-3: Publishing on the GEO Knowledge Hub. Lessons learnt from [EO-based surveillance of Mercury pollution](#) (Minamata Convention) Pilot - [Health Surveillance](#) showcase example.

The pilot is providing the workflow adopted in e-shape for the implementation of the Pilot [EO-based surveillance of mercury pollution](#) through the GKH (<https://gkhub.earthobservations.org/packages/2wxxd-w9009>).

The package provides information on how the tools of the GOS⁴M were built, the dataset used and provided, the scientific references, as well as presentations and videos. The GKH is a concrete opportunity to move from data sharing to knowledge sharing.

12.5.5 Publishing on NextGEOSS

12.5.5.1 NextGEOSS features basics

NextGEOSS is a contribution to EuroGEO which is a European contribution to GEOSS. It is the result of an H2020 Project which will end in November 2020 and its results will be sustained via e-shape, via the NEXT-EOS GEO Community activity and via other national projects. Publishing data or services in NextGEOSS is free.

Publishing in NextGEOSS can be considered faster and simpler than in GEOSS, especially for smaller data sets as NextGEOSS provides wizards to define static metadata that is compliant with NextGEOSS is already covering data, and services and proposes various offers:

- "[Link your Data](#)" - for those willing to see their data in the NextGEOSS data catalogue
 - OpenSearch harvester configuration - best for data catalogues exposing OpenSearch, dataset and data resources will be available
 - metadata form submission - the simplest way to make the datasets visible in the catalogue
 - custom harvester implementation - most flexible for large dynamic repositories where OpenSearch exposure is not used
- "[Showcase your Application](#)" - for those willing to promote their Services description and linkages in the NextGEOSS service catalogue
- "[Integrate your Pilot](#)" - custom option for deeper pilots integration, for example, cases where NextGEOSS data and services are combined.

The offer invites any data producer with a higher focus on the H2020 projects, to promote and increase the visibility of their data to a wider GEO Community. NextGEOSS advocates a resilient network of resources by creating alternative routes to making your data findable on a reliable basis. NextGEOSS resources are also findable via Google Data Search.

The data hub provides metadata about the origins of the data to credit the data providers as due.

All the data can be linked to concrete use cases to inform the visitors about what kind of use the data has been successfully used for and can faster and better evaluate if the data is what they need for their

use case. They will also be able to give their valuable feedback enabling the data providers to better target their user community, and improve their data or their communication.

The "Showcase your Application" offer invites any application developer with a higher focus on the H2020 projects, to promote and increase the visibility of their applications to a wider GEO Community. Cataloguing application or service results on NextGEOSS contributes to defragmenting the European contribution to GEO, leveraging European investments collectively supporting GEO and the global GEO community with findable new information, that can be applications or services.

12.5.5.2 NextGEOSS publication process

Publication of information of any kind requires one to create an account on the [registration page](#). Once registered one is able to create a custom OpenSearch harvester on the [catalog view](#) or service request through the [Service Desk request](#) of data or service link/catalog.

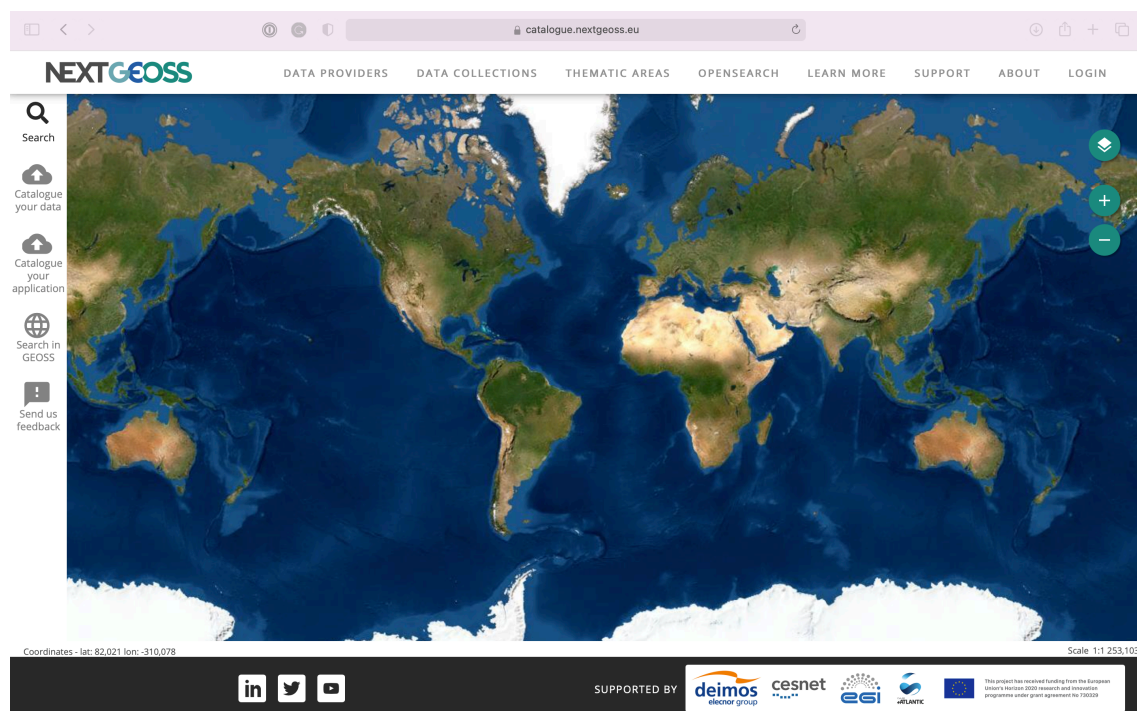
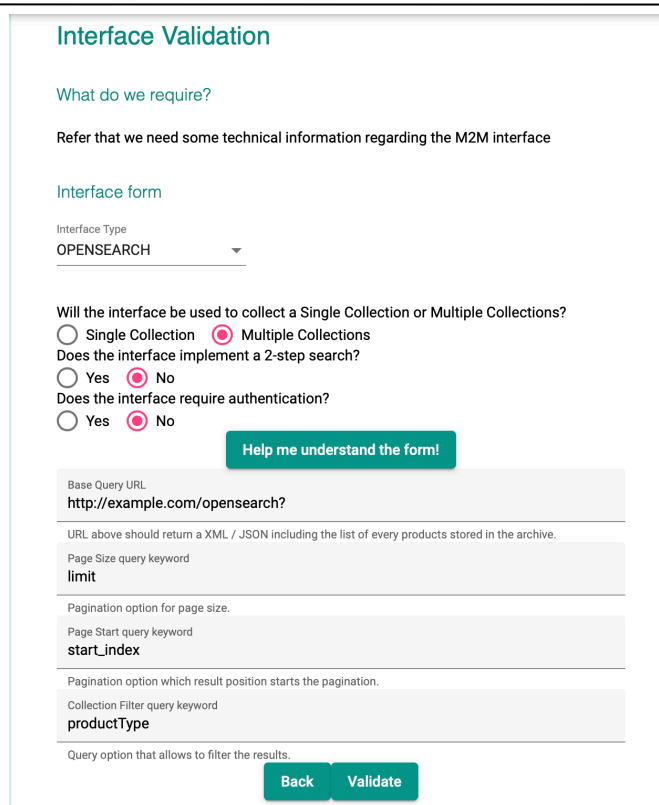


Figure 31: NextGEOSS data portal

For organisations exposing OpenSearch with Atom/JSON encoded metadata compliant with ISO19115/19139, it is recommended to rely on their own harvester. A successful configuration allows for a seamless integration of the services without much effort.



Interface Validation

What do we require?

Refer that we need some technical information regarding the M2M interface

Interface form

Interface Type
OPENSEARCH

Will the interface be used to collect a Single Collection or Multiple Collections?
☐ Single Collection ☒ Multiple Collections

Does the interface implement a 2-step search?
☐ Yes ☒ No

Does the interface require authentication?
☐ Yes ☒ No

[Help me understand the form!](#)

Base Query URL
http://example.com/opensearch?

URL above should return a XML / JSON including the list of every products stored in the archive.

Page Size query keyword
limit

Pagination option for page size.

Page Start query keyword
start_index

Pagination option which result position starts the pagination.

Collection Filter query keyword
productType

Query option that allows to filter the results.

[Back](#) [Validate](#)

Figure 32: Ticket registration over NextGEOSS Service Desk

The Service Desk requests cover all the other requests based on the issue categories. The ticket registration forms gather basic information about the dataset, which - depending on the data - can be registered directly into the catalogue (for singular resources, services), integrated based on the standard harvesters (OAI-PMH, CSW etc.) or into custom implementations. Service Desk forms encourage the submission of optional metadata files compliant with ISO19139. An online ISO19115/19139 generator tool is [available](#) where one can create sample ISO19115/19139 XML files simply by filling the web forms without any deeper knowledge about these standards.

12.5.5.3 NextGEOSS further information

- Manuals on how to publish data and services: <https://nextgeoss.eu/2021/05/07/how-to-catalogue-services-and-applications-on-nextgeoss-data-hub/>
- Webinar replay "Cataloguing Earth Observation in NextGEOSS" (19/5/2021): <https://event.webinarjam.com/replay/110/g3924s2ob6yf88uz9g>
- More on the "Link your data" offer: <https://nextgeoss.eu/join-us/link-your-data/>
- More on the "Showcase your Application" offer: <https://nextgeoss.eu/join-us/show-case-your-application/>
- Data and Service Cataloging: <https://nextgeoss.eu/wp-content/uploads/Data-and-Service-Cataloging-User-Guide.pdf>

12.5.5.4 Link your data

Currently : <https://nextgeoss.eu/wp-content/uploads/Data-and-Service-Cataloging-User-Guide.pdf>

12.5.6 Google Datasets

Dedicated Google Dataset search : <https://developers.google.com/search/docs/advanced/structured-data/dataset>

Mapping between the ISO-compliant CWS into the [schema.org](https://geo4web-testbed.github.io/topic4/#h.blzokpxksi4l) is <https://geo4web-testbed.github.io/topic4/#h.blzokpxksi4l>

12.5.7 Publishing on the DIASs

Each DIAS proposes a user guide addressing publishing:

- Onda showcase page provides a publication space for services developed/deployed on the Onda platform (<https://www.onda-dias.eu/cms/marketplace/cataloguem/>)
 - static presentation of companies/services and projects - self-managed, moderated publication form
- Mundi marketplace (<https://mundiwebservices.com/marketplace>)
 - static information about companies
- CreoDIAS proposes the following publication options for the partners and clients using the platform:
 - short organisation card on the <https://creodias.eu/partner-services> - Logo, title, ca. 1000 signs description, hyperlink.
 - data publication on <https://browser.creodias.eu> - static layer or data publication on the <https://finder.creodias.eu>
 - processors publication on the <https://finder.creodias.eu>
 - data publication on the dedicated mirror site like <https://cryo.land.copernicus.eu/finder/>

However, it looks like the latter 4 is for the moment not standardised even on the procedural level not saying about technical integration. However, they have experience and some implementations that support only selected standards.

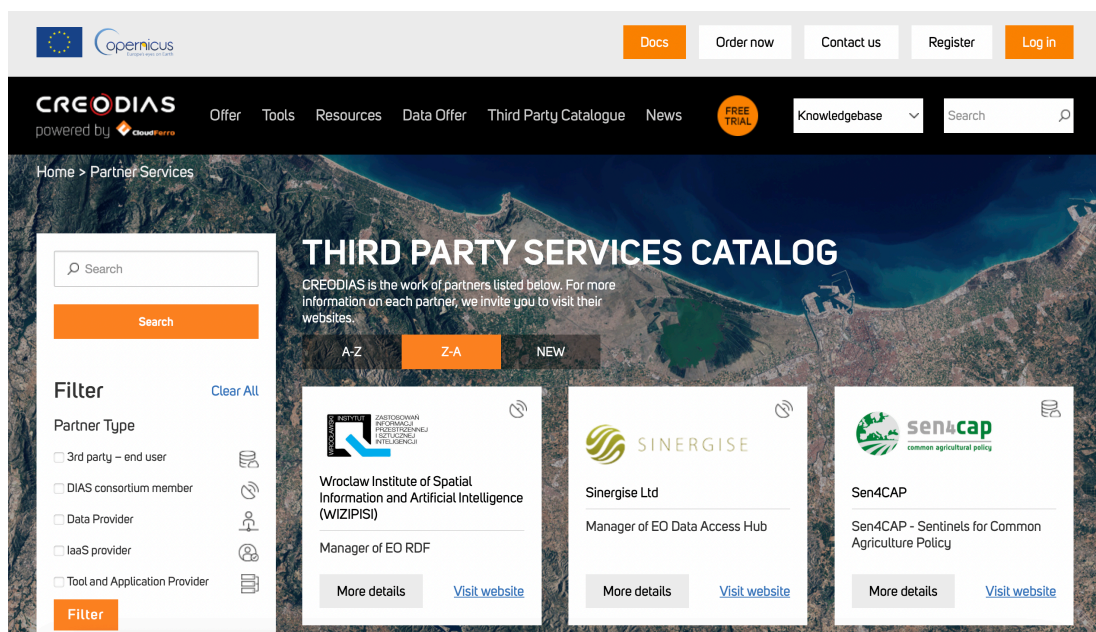


Figure 33: CREODIAS third party services catalog

12.5.8 Publishing on EoMall

EoMall is the platform provided by EARSC that supports the matchmaking between European EO market stakeholders. EoMall hosts knowledge granules on Earth Observation (EO) services. It is not designed for machine-to-machine (interoperable) information exchanges, but seeks to promote EO services to (non-EO) communities of professionals.

The market for EO services is undergoing a profound transformation along with the acquisition capabilities of both the public and the private sectors; it is moving from its traditional bespoke character to an online presence, where users can browse, access, and consume services from the various available platforms.

While this presents an enormous opportunity for both providers (who can expose their services to larger audiences) and users (who can select from a wider range of available services) it also brings significant challenges. The environment of EO platforms is rather complex with multiple, seemingly similar, vendor propositions, different metadata structures, functionalities, etc. It may therefore be hard for providers to know what the appropriate way to present their services is or for users to find what best suits their purposes.

The role of these platforms is all-the-more vital considering the globalization of the EO market and that 94% of the companies in the EO industry are micro or small companies. Platforms having developed software and tools can be a game-changer for solution providers struggling with IT limitations. At the same time, platforms can serve as marketing tools amongst users from different market segments; as the market has grown more specialized and segmented, limited time and capital saw solution providers focus their efforts on market segments they knew best.

Nevertheless, despite the intelligence and tools made available via these platforms, promoting their use remains a challenge. A study made in 2018 by EARSC and BHO Legal concluded that technically advanced potential users are aware of these platforms, but few use them claiming:

- Difficulty finding them (42% of respondents),
- Lack of transparency regarding offer terms of use (58% of respondents considered that accessing the terms and conditions of EO platforms prior registration is not straightforward and offers are not immediately apparent),
- Challenging user interfaces (83% of respondents considered EO platforms to be difficult to navigate)

To follow this sector market evolution, the European Association of Remote Sensing Companies (EARSC) has developed two complementary online resources eoMALL and eoWIKI, which represent the digital tool kits to promote EO services online.

eoMALL hosts 52 entities in total, of which 34 EO and currently 15 are concretely using this platform. 140 registered users, 39 companies have signed the Chart.

Of the 34 EO companies displayed on eoMALL, 14 countries are represented with Netherlands, Germany and Spain respectively 6 and 5 companies being the top 3 countries followed by Spain (5), France (3), Italy (3) and Greece (3). At the moment, 10 companies are e-shape Pilots.

In terms of visitors, the main come from USA, China and UK; of the eoMALL for EuroGEO page, the three mains countries are Poland, France and Belgium.

Testimonials from e-shape Pilot promoted through EO MAIL

Planetek's EO service Rheticus refers to the platform as "Planetek's cloud-based EO services based on Rheticus are displayed on eoMALL which is a great advantage as this platform allows to compare different EO services that can best suit the users' demand, as well as support uptake and awareness through eoWIKI success stories".

NASU-SSAU, the Ukraine Space Research Institute, makes use of the eoMALL for EuroGEO page and “From a research-to-business perspective, very interesting section within this General Assembly”.

eoMALL is seen as a "window-to-the-market" platform to showcase EO services to a broader plateau of users and reaching to them outside the traditional fora of conferences and events, which, of course, are the main places to promote online tools. The platform is introduced also during dedicated meetings that EARSC holds for the institutional liaison, with the association of categories representing final non-EO users as a concrete example of how the e-shape pilots support the dissemination of their services among different communities.

12.5.9 Publishing on EoWiki (knowledge component)

eoWIKI is an accessible and open-to-all online platform resource, it has been connected to the e-shape project to sustain the acquisition and knowledge transfer, facilitating the user-provider interactions by delivering key information through best practices, success stories, news on technology, and market trends.

Currently, EARSC has created and populated 17 success stories supplied by e-shape pilots. Information was collected in an eoWIKI template. These success stories are embedded in the success stories main page, together with success stories coming from other projects, and accessible through the search engine or under the taxonomy thematic sectors.

eoWIKI also hosts an “eoWIKI for e-shape” page, dedicated to promoting the engagement and the links of the pilots with the broader ecosystem of communities of users. In this way, the web user can:

- Navigate and identify how EO is used through a variety of cases (EO taxonomy)
- Discover the e-shape success stories, enhancing the visibility of the e-shape pilots' success stories)
- Get insights into user-related challenges (best practices)
- Discover market and technology trends, and business guidance through the e-shape sustainability booster (sustainability)

12.6 Making data accessible via FTP (File Transfer Protocol)

FTP stands for File Transfer Protocol. It is a long-standing method that allows copying files from machine to machine. Tools such as Filezilla can help publish and access data. It can be secured with Kerberos for instance to manage the authentication and encryption of the data. It is convenient when you need to transfer large amounts of data. If data needs to be processed on the fly, APIs will be more suited. FTP is often used in backup architectures to secure data access.

12.7 Disseminating data via Satellite dissemination

Data can be disseminated (pushed) via ground network and/or via satellite dissemination. Under the footprint of the satellite dissemination, the reception is done via a satellite antenna that is usually quite inexpensive.

GEONETCast is a global network of satellite-based data dissemination systems providing environmental data to users around the world. It is part of the GEO System of Systems - GEOSS. It gathers several broadcast streams: EUNETCast, CMACast and GEONETCast.

As these dissemination channels are inexpensive, timely, and very operational, they are also very intensively used and it can be difficult to add a new product to the dissemination catalogue.

Lessons learned on EumetCAST:

- EumetCAST is a very reliable service and very valuable source of data(see above “experiences are perfect”) to provide data over the EumetCAST footprint
- EumetCAST provides mainly environmental data
- It has more than 4000 customers that are probably the major stakeholders in its footprint.

Delivery Media	Description
<u>EUMETCast</u>	EUMETCast is the EUMETSAT contribution to GEONETCast with coverage over Europe, Africa, and the Americas. Established in 2004, EUMETCast has more than 4,000 registered reception stations with more than 3,200 users benefiting from the environmental data it provides.
<u>CMACast</u>	CMACast is the China Meteorological Administration’s contribution to GEONETCast. CMACast utilises the AsiaSat 4 satellite beam to broadcast data and products to a user community in the Asia Pacific region.
<u>GEONETCast Americas</u>	Broadcast covering the Americas, managed by NOAA.

Box 12-4: GeoNetcast. Lessons learnt from the agriculture VICI - [Vegetation-Index Crop-Insurance in Ethiopia pilot](#)

GeoNetcast (see <https://www.earthobservations.org/geonetcast.php>) is managed/operated collaboratively by China (CMA), EUMETSAT, and the US (NOAA). Eumetsat greatly facilitates the timely and fast distribution of trillions of image data to end-users, located anywhere, at near-to-zero costs. In locations where data volume, download speed, and internet speed/reliability are limited, GeoNetcast offers the (only) way to remain in business. GeoNetcast is specifically used for the distribution of meteorological data (globally). Experiences are perfect.

Partners in Mekelle University have experienced the use of GeoNetcast as their only means to obtain RS-Imagery. To date, e.g. we captured on a dekad basis, all Proba-V 10-day NDVI images (Africa-window) through GeoNetcast. Local processing of many captured images takes place through the Ilwis platform (software). For VICI, specific routines as required for insurance purposes were developed, and are operational both at the NMA in Addis, ad Mekelle Uni in Tigray.

Box 12-5: EUMETCast. Lessons learnt from Pilot2: [High PV penetration in urban area](#) (economic opportunities for solar energy through urban solar mapping) of the [Renewable Energy](#) showcase

The Pilot relies on Near Real-Time Meteosat Second Generation satellite, SEVIRI HRIT Level 1.5 observations disseminated via Eumetcast as input data.

12.8 Publication standards

As visible on multiple platforms, data publication and dissemination activities can reuse the same interoperability mechanism both between platforms and various resources.

Once exposed, the standardised interface can be consumed by a number of client applications making data accessible widely. As the standard-compliant metadata preparation is the issue in itself and can consume significant time if done for the first time and manually, it is recommended to use metadata generators (See NextGEOSS publication process chapter) for static descriptions or tools exposing standardised interfaces like GeoNetwork, CKAN, PyCWS (see Data and Service metadata chapter). Nevertheless, it is useful to understand what are the similarities and differences between standards.

1. ISO 19115 - the international standard that defines what information should exist in geospatial metadata without specific encoding constraints
2. ISO 19139 - produces an XML Schema defining how metadata conforming to ISO 19115 and 19119 should be stored in XML format.
ISO 19139 requires that the encoding implements one of the following standards:
 - [CSW2 AP ISO] XML Schema13 - <https://www.ogc.org/standards/cat>
 - [ISO 19139] XML Schema as available in the ISO repository 14, or
 - [ISO 19139] XML Schema as available in the OGC schema repository 15.
3. CSW - OGC Catalogue Service (formerly 'for the Web') is the standard way the metadata can be exposed for querying and discovery. It is ISO 19115/19119/19139 compliant.
4. Dublin core - a defined minimum set of the metadata required for proper resource description (like ISO 19115, but simpler and defined on the conceptual level)
5. INSPIRE EUC directive refers to the above standards while defining the minimum metadata set required for EU resources. See "Technical Guidance for the implementation of INSPIRE dataset and service metadata based on ISO/TS 19139:2007" - <https://inspire.ec.europa.eu/id/document/tg/metadata-iso19139> for a detailed comparison between metadata constraints.

The following chapter describes what information metadata shall carry to satisfy all the above information models and be supported by their encodings.

12.8.1 Data description best practice

The following considerations provide the minimal content to create a metadata record. Not all the questions may apply to the product you are trying to describe.

1. Id - globally unique identifier of the resource
2. Title - simple while comprehensive and unambiguous name of the resource, catalogue or service
3. keywords - keywords are both the common terms that characterise the resource enabling finding similar ones and those that differentiate it from the other similar (keywords shall not be temporal nor spatial extent and other code-listed parameters included in the metadata). Use code listed or taxonomy-based words.
4. Abstract/Description - an abstract form precise description of the resource written in a common language, without ambiguous abbreviations (description shall limit repeating already structured metadata data like temporal and spatial extent, other code-listed parameters to a minimum)
 - a. Purpose - Why were the data collected and how is/will be used or how could they be used by future researchers
5. Spatial extend - what is the approximate location of your data or applicability of the service? These can be county names, states, general regions, NPS units, lat-longs bounding coordinates, and place keywords with an unambiguous reference system. For catalogues, It is important to define the minimum extent of the actually available data, not potential.
6. Temporal extent - time period represented by the data. Similarly to a spatial extent, it shall be an actual min-max time relevant to the resources described.
 - a. product generation time is often required (see ISO) if significantly different from the observation extent

-
7. Resolution - spatial and temporal resolution shall be described for each data resource and if possible for the dataset as min-max
 - a. if various, spatial resolution shall be defined in each dimension
 8. Originator - organisation and person identified as the owner and contact point for the data set (who created the data and who is responsible for the data available)
 - a. (list of) name, address, telephone number, and email of Points of Contact
 - b. The organisation that is responsible for the data and service, and any other organisations with significant contributions that shall be credited.
 - i. formal name, address, web page
 9. Data address with formats - shapefile, raster, spreadsheet, database, ArcInfo coverage, text file, other (please identify), specific standardised formats and generating software version is encouraged
 - a. if the resource is available in various formats via various endpoints, all shall be listed with format/interface information
 - a. Legal reusability - Are the data sensitive or classified or are there any legal restrictions on who may obtain/use the data?
 - b. The license shall be linked to the data to enable reuse, but general information is useful for the general community at the start.
 10. References - are there any publications associated with this resource or related works that help explain the methods or content of the data. Any source datasets used for processing, validation, etc that will help understand the data.
 11. Status - What is the status of the data you are documenting? – “complete”, “in progress”, or “planned” (Status)
 - a. update or new products in catalogue frequency Weekly, Monthly, Annually, Irregularly, or As Needed
 12. Taxonomic information - names, ontologies
 13. Methodology - Well-known, established/published methods or techniques of data production or analysis work. If you used standard, published protocols/methods, simply put the complete citation for the reference in references. Essentially, if a method is well documented it can simply be referred to and listed with a citation, rather than in detail as a ‘processing step’
 - a. existing protocols or methods
 - b. data processing model or other analytical tool, URL/contact if available
 - c. measures (if any) to make certain that your data set was as correct as possible (e.g. instrument calibrations, spot-checking data, spreadsheet macros for outliers, accuracy assessment matrices, etc.)
 - d. things excluded from your data collection like specific objects, confidential data, time periods
 14. Optional - data description reference; ideally external resource with:
 - a. short description of what each field in the attribute table means (include units of measure if applicable). This section will describe the fields and values of the dataset. (Entity & Attribute). Make sure future data users will understand what the fields represent and the definitions of any values that they contain.



-
- b. Do any values in the dataset represent codes from a data dictionary or code book (Taxonomic or biological abbreviations, etc.)? If so, please provide references for where these values can be explained. This can be a more efficient way to document for future users what values mean, instead of providing full, detailed explanations within the metadata document itself (Entity & Attribute – “Codeset”
 - c. List data sets used to produce this data with URI (if available) and catalogue or source name, originator and publication date, time period and geographic scale, source presentation form and media type, and contribution of the source to your analysis
 - d. Any advice for potential users of the data set

Section inspired by the "U.S. Geological Survey - Data Management Guidance Materials / White Papers" (Various Authors) Resources obtained from USGS staff and/or the USGS Data Management website (<https://www.usgs.gov/datamanagement>), November 2015. Documents may be subject to revision.

13 EXPLOITATION

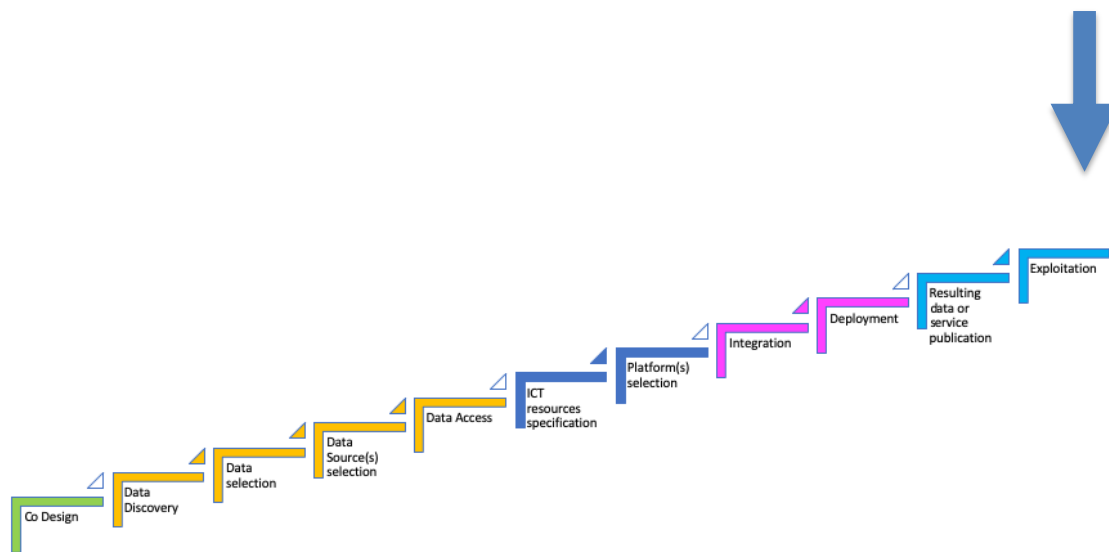


Figure 34: Exploitation in the Development workflow

13.1 Common terminology ambiguities

Some words such as "Real-time" or "Operational" are commonly used and carry a lot of ambiguities. These concepts are relevant to most EO use cases and have to be explained clearly in any new working context, new partner interaction, or any legal agreement to avoid any later problems.

Use cases are usually "real-time" or over past data. Real-time refers to the ability of a system or process to respond to an event or request immediately as it occurs, without any significant delay. In the industry, this can be milliseconds, and in Earth Observation it will most of the time be seconds to a few minutes for data needed for crisis situations, and eventually some hours for numerical models for instance. The definition refers to short-term use. It can be linked to the scientific temporal validity of the data, to the timeliness of the usage scenario, or to the technological capacities. It is always useful to qualify the timelines: is it some seconds requiring intense validation of the communication or processing capacities? Is it some minutes, one hour to have time to apply some basic data consistency controls? or can the need wait for several hours to have a better quality dataset or product where more sophisticated quality controls have been applied? Sometimes, the need in term of timeliness can lead to define several products with different latency/quality level combinations.

Then, beside real time uses, Earth Observation Data is always potentially useful in the long term for climate study impacts.

"Operational" refers to a system that can carry out an intended purpose without impacting disruption. This means that a platform or system offering a service can be operational because it provides sophisticated capacities and can be used for many productions and suffer interruptions of several hours for maintenance or other without impacting its "operational" qualification if this is acceptable for the user. This is not the case for an emergency crisis situation, for businesses, or for applications on the web or on mobile. In this case, the mean time between failure, the recovery delay, the process in case of a problem, and the performance of the support will be critical. They have to be considered before subcontracting a service and will have to be discussed in the contract design process.

Box 13-1: Exploitation. Lessons learnt from the Pilot [Diver Information on Visibility in Europe](#) (coastal water quality monitoring) of the [Water resources management](#) showcase.

The Pilot developed a web application deployed on google store relying on a DIAS. The DIAs suffered from disruptions in data updates and the support has not been reactive enough compared to the operational need. As a matter of fact, users who install an app that does not work, deinstall the app rapidly and do not come back. Moreover, they often put feedback on the store that will prevent new visitors. This resulted in the need to change the data source under pressure and a big feeling of disappointment from the developer. The same DIAS was appreciated positively by other e-shape pilots, meaning that it was able to provide an "operational" service but not for mobile apps requiring frequent data updates.

13.2 Describing the fitness for use and limits of operational products

Products are often advertised based on their benefits and can perform the tasks or functions it was designed for effectively and efficiently but they can have limitations in their fit for use, i.e. their suitability for an intended purpose. These limitations should be described in an exploitation document that should be short, efficient, and considered with each operational product.

Box 13-2: Fitness for use. Lessons learnt from [Assessing Geo-hazard Vulnerability of Cities and Critical Infrastructures](#) Pilot [Disasters Resilience](#) Showcase Example

SAR data and derived products are useful to detect and monitor slow-moving geological hazards like land subsidence, long-term landslides, building settlement, and mining activities. In the case of landslides, detection is possible when the displacement rate is below 40 cm/year. For slower case studies, analysis of the InSAR products is possible and reveals the differential displacements in the landslide. Also, potential damage to structures and infrastructures can be achieved. A sudden collapse of the phenomena could be estimated if accelerations are detected in the Time Series. These precise results are obtained after the analysis of data and correlation with in-situ and auxiliary information. On the other hand, fast and sudden events can only be detected after the collapse and using other interferometric products. In terms of time response, interferometric products are difficult to process and analyse. First results obtained by automatic tools like GEP-TEP can be achieved a few days after an emergency but only the most significant features can be detected.

Regarding floods, interferometry products are capable to detect the extent of floods using the specular response of radar waves over water. Even SAR Amplitude and Coherence are useful to provide this information, those products are not the main focus of the pilot which takes advantage of Interferometric products. Moreover, this kind of fast response to flood events is already provided by Copernicus Emergency Management Service.

13.3 Managing input data changes

Satellites, Sensors, and Production processes have a lifetime duration, or the conditions to their accessibility can change and impact the downstream business model for the users. In the case of perennial missions such as geostationary Meteorological satellites, renewed each 15 to 20 years, the data users are informed well in advance and can access prototype data before the change of satellite to prepare their continuity of service to their own customers. It is nevertheless a tremendous work to compare the input data, adapt the algorithms, compare results, and recalibrate. This tackles the issues of continuity of products and reusability of the applications using this data as input data (the "R" of the FAIR principles). These issues are probably even more critical in the case of less operational and well-established communities, as they might be informed much later, and have more difficulties accessing prototype data of the future replacing data or products.

Continuity of products is important for downstream production but also for temporal consistency in time series. This is a real challenge as the different satellite sensors have differences in radiometric calibration, differences in spectral bands, and differences in orbital characteristics, and scanning systems.

This can sometimes be the opportunity to replace the historical products by an improved product, based on an improved algorithm but this might require updating all the past references with the same improved algorithm. This is called historical data curation. This can require a lot of resources and time depending on the historical depth and the data volumes to be reanalysed. Then, if all the downstream users cannot access this new product or adapt their production chain to this new product at the same time, this will require double dissemination or a very well-coordinated migration plan impacting all the downstream users.

In a summary, to fully manage the operational impacts of the end of satellite products, it is necessary:

- For the data provider:
 - to inform the downstream users about the end of the product availability,
 - to inform the downstream users about the new alternative potential products with details on the differences between the old and new products,
 - to provide well in advance some simulated data,
 - to disseminate the new alternative product(s) via the same channels as the old ones or give access to the new alternative product(s) at the same endpoint. This can be hard when dealing with satellite broadcasts, with usually very busy bandwidth and used to send data in areas where the terrestrial networks are poor and cannot be a reliable alternative solution.
 - To inform as early as possible of any change related to the protocols or API to access the data.
- For the user of the data product:
 - to test the alternative product(s),
 - adapt his downstream processing to the new input data,
 - to test the access to the new data,
 - Eventually recalculate past data based on the new downstream processing or consolidate time series based on ancient and new inputs.

The end of life of satellite products has to be carefully taken into account when defining a new contract to deliver any downstream service.

Box 13-3: Managing input data changes e-shape agriculture VICI - [Vegetation-Index Crop-Insurance in Ethiopia](#) pilot experience.

The e-shape pilot VICI - [Vegetation-Index Crop-Insurance in Ethiopia](#) provided a financial service delivery to smallholder farmers in Ethiopia with a geodata-driven risk-mitigation (insurance) product that offers a basic safety net to protect them against weather-related perils. The processing chain to deliver NDVI products was based on PROBA-V Copernicus NDVI and SPOT-VEGETATION products since May 2013. Due to the retirement of Proba-V satellite planned in June 2020, the PROBA-V products were planned to be replaced by OLCI-SLSTR Synergy (Sentinel-3) comparable products. The first challenge for the pilot, was to work on the continuation of PROBA-V Product with Sentinel-3 Synergy 10 Product for Vegetation to prepare operational migration of the production before June 2020. The expected OLCI-SLSTR Synergy (Sentinel-3) production has being delayed, the pilot has found a good substitute solution based on Copernicus CLS BRDF corrected also based on Sentinel 3 with a 1 km resolution that is available to the public since July 2020.. Moreover, a data archive for this product is available for over 20 years and there are plans to deliver 300m resolution products in the future.

The PROBA-V data was provided via GeoNetCast whose existing band-width is already under pressure. Adding the BRDF products to these broadcasts requires support and time.

The pilot then had to freeze certain pilot activities because they needed to first know if they would have an index insurance product. As it happened, COVID-19 prevented any field work from happening and they had to wait for better conditions.

Collaboration between the pilot and the JICA ICIP project in Ethiopia remained excellent. ICIP has been doing the marketing, fieldwork, and implementation of the insurance in one region of Ethiopia (Oromia) in 2020 and has ensured continuity for four more years - IF - the pilot could generate and maintain the product.

With specific thanks to VITO, they were able to receive copies of the past 20 years BRDF-adjusted NDVI-images (of Spot-VGT and Proba-V; 1km resolution). These images were fully processed to recalibrate VICI and make it ready for data input changes (from Proba-V to Sentinel-3). For a while, however, that switch could not be implemented, and the pilot remained dependent on the Proba-V images (the processing line was re-activated by VITO).

Contractually, ACRI was in charge to produce 1km Sentinel-3-based NDVI-products. After many delays, and suggested adjustments, ACRI managed by 1 June to produce the first NDVI-images (not BRDF-adjusted). These images lacked proper radiometric adjustment and in many places (likely cloudy areas in Ethiopia), reported NDVI-values went from way too low to way too high (with no means to post-process them properly).

In the meantime, VITO (contractually) was preparing proper BRDF-adjusted NDVI-images at a 300m resolution, based on Sentinel-3 data. Although prepared, these images were still not in circulation (pending admin approvals), Which required VITO to re-open their Proba-V processing line, knowingly that the Proba platform was seriously drifting (overpass was earlier and earlier).

To adjust data captured by different platforms that have different (daily) overpass time as to adjust differences in solar angles (across days within a year), which all have impacts on the shade etc., BRDF adjustments is widely recommended. That adjustment can only be implemented by the data providers.

For a while still, the 300m Sentinel-3 based BRDF-adjusted NDVI-products had NOT been made available. Regularly, additional delays happen (admin./institutional at EU-level).

Once these data are disseminated, the pilot strongly requested that these 300m Sentinel-3 images must be distributed through GeoNetCast (request for Eumetsat). Many institutes/organizations, specifically in Africa, can only depend on that specific medium. This request is fully supported by JRC.

- the VICI insurance, Ethiopia, fully depends on the above.
- See video: https://dikke.itc.utwente.nl:5001/fsdownload/TnLUMyC6W/DIAS_eShape-S1P3_VICI#

13.4 Single point of failure analysis

Being operational can have different meanings in different communities. Some information systems are critical for safety or business and require high availability and reliability. These information systems will require a Single point of failure analysis. A Single Point of Failure (SPOF) is part of a system, that, if it fails will stop the entire system. This can be data, software application, infrastructure, network, or any system contributing to the production process.

The Single point of failure analysis is a process of reviewing, identifying, and evaluating the potential risk systematically for each data source, piece of software, infrastructure, and network. The goal of such an analysis is to identify the potential risks associated with these critical components and develop a plan to mitigate those risks or their impacts.

Box 13-4: Single point of failure analysis. Lessons learnt from The Pilot 2: [GEOSS for Disasters in Urban Environment](#) (improved resilience of cities, infrastructure and ecosystems to disasters) of the Showcase Disasters Resilience

The Pilot developed activities related to the execution of a hydro-meteorological forecasting chain including the WRF model, the hydrological model Continuum. The WRF model assimilates different data sources provided for research purposes, namely the radar data of the Italian Civil Protection Department (ICPD) mosaic, the weather stations of the ICPD and the personal weather stations data provided by the Wunderground network. All those data are available in real-time mode and operational manner, since CIMA Research Foundation is part of the Italian National Civil Protection System Network of Excellence within the fields of floods and forest fires, with a special focus on early warning systems.

Failure, albeit very uncommon, of data streams may happen. The pilot has explored the possible usage of different combinations of data to be assimilated to mitigate the risks of such failures. In this case, the analysis will assess the impact of the disruption of some data sources to assess the most critical ones and if the diversity of data and sources provides sufficient resilience to guarantee acceptable quality to the outputs at any time.

13.5 Using Web Analytics tools or services to optimize the publication

Usually, products are co-designed with users in face-to-face meetings and validated by the primary users. Developers know the users. After a web publication via one to several portals, user interfaces, search engines, links in documents or social media will support the outreach and upscale of the product's usage. At this stage, the developers don't know the users anymore. The only way to learn about the new users will be via web analytics services and tools. Web Analytics tools provide data and insights on website usage and visitor behaviour; This data can provide valuable insights into how users interact with a website, what pages are most popular, where users are located, and much more. They allow one to track and analyse user behaviour on a website to optimize the website for the target audience. There are many analytics tools available in the market such as Matomo, Mixpanel, Hotjar, Woopra... They can be compared to Google Analytics which is the most popular but suffers some critics related to GDPR.

These tools allow us to know the number of visitors at any time, the number of new visitors or returning visitors, the number of visitors per country, the language of the visitors, the devices they are using (i.e. desktop, mobile, or tablet), the acquisition channels (i.e. if the users reached your website from a social media, a direct link, an organic search...), which keyword they used for their search, the number of times a page is viewed, the time visitors spend on the site, ...

Without knowing personally his users, the developer can discover where he can expand his market, in which language it is worth translating the website, what keyword he should emphasize in his website because they are the most searched, what data is the most searched, accessed and retrieved, which data is never accessed revealing either a lack of interest for this data or a problem in its publication made with inappropriate keywords. The developer is then able to consolidate its publication, emphasize the products that are the most popular, rank the search results using these popularity criteria to help the user find more easily the data of interest for them, and for instance optimize its online/offline catalogues to make the most searched data more easily accessible. He can communicate about the products that looked to be underused.

Of course, the data needs to have a minimum audience to collect significant Web Data analytics and there can be a syndrome of the chicken and the egg in the initial phase after the publication.

These Web Data analytics can be the foundation for Technical and Marketing strategy as well as for a complete business model. They reveal the societal and business value of the data or products, being these data open or commercial. They also reveal the "usability" of your system, the effectiveness of the discoverability, the relevance of your metadata, the potential new markets and so much more. It is a tool itself to monitor daily the good reliability of your system and to forecast in some cases its problems. They can reveal problems linked to the number of visitors or to a rapid increase of the audience for instance. They can be used by the operational teams as a system monitoring tool.

The use of these data analytics should be developed as a powerful tool to upscale. At the time being, they are underused and there are concerns that some GAFAs make better use of them than the real data producers.

Forestry conditions (more efficient forestry operations with lower environmental impact and carbon emissions) Pilot service from Climate Showcase Example

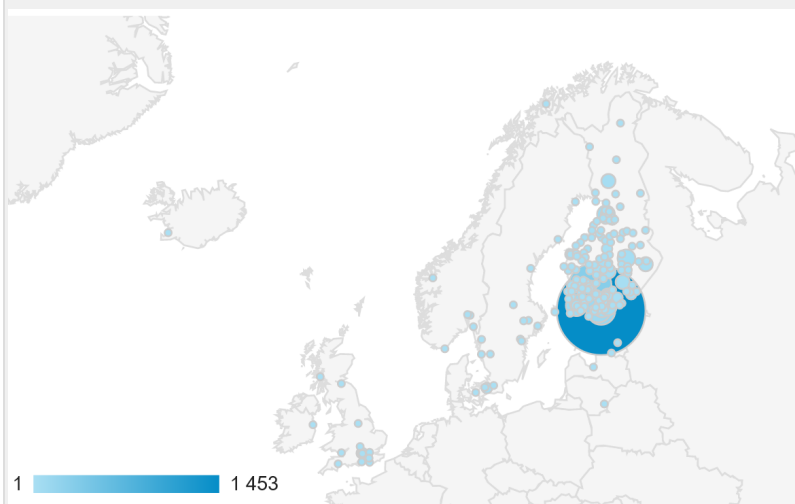


Figure 35 HarvesterSeasons.com usage as analysed by Google Analytics

HarvesterSeasons.com usage as analysed by Google Analytics shows that sage is well spread all across forested Finland.

Box 13-5: Web analytics. Lessons learnt from [the Forestry condition](#) Pilot.

The Pilot Forestry conditions (more efficient forestry operations with lower environmental impact and carbon emissions) of the Climate showcase has developed a service supporting the forestry sector: harvesterseasons.com. It has disseminated service promotion via several platforms and portals. The service has been using Google Analytics in the background to analyse the impact of all dissemination activities respectively. It is important to state that the Analytics service requires a statement in the privacy policy, in which user information is collected for statistical analysis. In the case of this pilot, it is only basic information on location, IP, and acquisition channel. But even with only basic information Google Analytics provides a good understanding of users, service usage, and the impact of dissemination activities.

Throughout the project timeline, it could be clearly seen that each promotion activity has impacted in a rise in service usage. Most importantly the service is findable directly and prominently via Google search. Additionally, to just name a few promotion activities, S7P3 had the Climate showcase webinar series, individual service webinar, participation at user events for the Finnish forestry sector, and presentation at conferences. Each of these activities helped to raise awareness of the forestry service and was directly visible in click numbers.

Through Google Analytics it was also possible to investigate various user acquisition channels. Pilot S7P3 has been promoted through various channels. Harvester Seasons' own LinkedIn and Youtube profile, proving monthly service updates and information, helped most effectively to bring users to the service platform. Articles at the portals of the Finnish Meteorological Institute, Copernicus, E-shape, as well as WEkEO DIAS (Harvester Seasons is one of the WEkEO use cases), are as well clearly visible as strong user acquisition channels according to Google Analytics. Last but not least the Harvester Season pilot is findable via GEOSS as well as the national GEOdata portal <https://kartta.paikkatietoikkuna.fi/>. Even though those platforms are not yet well known in the forestry community it could be shown via Google Analytics that the Harvester Seasons service acquainted at least some visits via those platforms.

Lessons learned on Web analytics

- The open data value can be revealed by the use of Web Analytics tools.
- These data are an opportunity to optimize the catalogues.
- Unfortunately, the most current free tools are US and their data is not open.
- As these data reveal the marketing value of open data, after some first analysis with open tools, developers should consider moving to paying web analytic tools to preserve these data confidentiality.
- The use of these data analytics should be developed as a powerful tool to upscale. At the time being, they are underused and there are concerns that some GAFAs make better use of them than the real data producer.

13.6 Reproducibility

Reproducibility for Earth Observation applications refers to the ability to reproduce the results of EO analyses or experiments by using the same or similar data, methods, and workflows. Reproducibility is essential for ensuring the reliability and accuracy of EO analyses, as well as for facilitating collaboration and knowledge sharing among researchers and practitioners.

Reproducibility in EO applications requires several key components, including:

- Open data, freely available, properly documented, and easy to access.

- Transparent and documented reproducible workflows used for data processing and analysis so that others can understand how the results were obtained. This includes documenting software versions, parameter settings, and scripts used for data processing.
- Shared code is available to others so that they can reproduce the results.
- Proper, clear, complete, and accessible. documentation of data sources, processing methods, and analysis workflows
- Version control tools such as Git for ensuring that data, code, and documentation are properly tracked, versioned, and shared.

Reproducibility in EO applications is essential for promoting transparency, accountability, and scientific rigor. It enables researchers and practitioners to validate and compare results, facilitate collaboration, and build upon existing knowledge to advance the field of EO.

Reproducibility is not only useful for research scientific evaluation and verification but it is also needed to reuse software and applications over other areas than the ones they were initially developed for to expand a market for instance. The developers will first run the application on the usual domain the application was designed for and when all the processes will be mastered, they will start adapting the application to a new domain, eventually changing the input data sources and verifying progressively the impacts of the changes and finally the results.

In order to go towards knowledge sharing and reproducibility, the e-shape project has supported the collection of knowledge packages that are now accessible to all via the GEO knowledge hub. Depending on what the pilots have provided, they can include open source, scientific papers, videos...

e-shape knowledge packages on the GEO knowledge hub are accessible at: <https://tinyurl.com/GEO-Knowledge-Hub>

13.7 Data Management Plan

Depending on the development context and goals, the data or application developed by an EO developer can have to comply with different legal frameworks or data policies such as the INSPIRE Directive, the GEO Data Sharing and Management Principles, the FAIR principles, the TRUST Principles, the CARE framework, the Regulation (EU) 2016/6791, the European Union's ('EU') new General Data Protection Regulation ('GDPR'), ...

These frameworks and policies will impact the development process, the resulting data or services management, and the operations.

The applicable regulations have to be identified, their impacts have to be analysed and integrated into the architecture, the developments or exploitation organization and the good compliance has to be verified in the best way. The earlier the better to initiate this process that will have to be reviewed and consolidated several times during the development process period of time.

Since the e-shape project is participating to the Pilot on Open Research Data (ORD) in Horizon 2020, it was mandatory that the e-shape consortium prepare a Data Management Plan (DMP). The first version of the e-shape DMP showed a deficit of attention on the GEO and FAIR data management principles, which should then be progressively promoted.

No canvas able to capture levels of compliance to the GEO and FAIR dimensions was available and a new specific canvas was required to capture in a homogeneous way over the many pilots, the level of compliance to the Data Management Principles for each of the e-shape pilot, in view of monitoring progress towards the achievement of the corresponding e-shape KPIs.

e-shape participants acknowledged the added value of conducting the DMP process, as it triggered a number of internal discussions and clarifications, which otherwise would not have been clarified in usual

practices. It was also acknowledged that the exercise had value in itself for the overall adoption and promotion of the GEO and FAIR principles across the EO community; by turning principles into an actual operational questionnaire. On a more practical aspect, numerous points raised by the GEO and FAIR frameworks as part of the DMP tool are not easy to address. These points address the standard format or protocol supporting the input as well as the output data and the process which generated such data. This was promoted as an opportunity for teams gathered around a pilot to internally brainstorm about aspects sometimes taken for granted (e.g. policies, procedures, processes, supported standards, and licenses) regarding input data, to encompass more prospective aspects when it comes to the concerns defining the output data and the processes that generated them. The DMP framework, and especially the GEO and FAIR principles addressed by the e-shape DMP tool, is a powerful framework for such brainstorming.

The FAIR and GEO principles are mainly built to support the notion of data and metadata. Nowadays the notion of service or “as a service” is increasing in the EO sector. The current granular questions around the GEO and FAIR principles does not address as such this notion of service or “as a service”. There is for sure room for improvement to make more obvious how these principles refer differently to data or services. A new GEO DMP review has been validated by the GEO Program Board in June 2022 introducing the notion of services and APIs as the previous release of the document only referred to data.

One issue it had to address, was the overlapping scope, yet complementarity, between the FAIR and GEO data management principles, which might have involved collecting redundant information. Also, the scope of GEO is perceived to be “Data-Centric” vs. e-shape which is “Service-oriented”.

The e-shape toolbox has been adopted by the 37 Pilots. It was endorsed by the GEO Secretariat and has been uploaded to the GEO Knowledge Hub, and is available as a DMP self-assessment tool: <https://gkhub.earthobservations.org/records/rtdy9-qnd28>

It is a Best Practice recommended by e-shape for the EO applications developer to use the tool to self-assess the Compliance status in relation to the GEO DMPs and the FAIR Principles.

It is also a best practice recommended by the e-shape project, to use the tool for projects implementing several pilots, to encourage and support progress on this compliance between the start and the end of the project, based on the notion of “trajectory”.

The toolbox was presented through the GEO dialog series:

<https://gkhub.earthobservations.org/records/kbgd8-58w14>

Lessons learned

- The initial level of familiarity with the GEO and FAIR Data Management Principles was low (37%), but targeted capacity building significantly improved the metrics (69%). Some GEO and FAIR dimensions are still underrepresented, which would require a more granular approach to education on specific aspects, possibly applicable to the larger EO community.
- The exercise had value in itself for the overall adoption and promotion of the GEO and FAIR principles across the EO community; by turning principles into an actual operational questionnaire.
- On a more practical aspect, numerous points raised by the GEO and FAIR frameworks as part of the DMP tool are not easy to address. These points address the standard format or protocol supporting an input as well as the output data and the process which generated such data.
- This should not be viewed as a hurdle but more as an opportunity for teams gathered around a pilot to internally brainstorm about aspects sometimes taken for granted (e.g. policies, procedures, processes, supported standards, and licenses) regarding input data, to encompass more prospective aspects when it comes to the concerns defining the output data and the processes that generated

them. The DMP framework, and especially the GEO and FAIR principles addressed by the e-shape DMP tool, is a powerful framework for such brainstorming.

- The FAIR and GEO principles are mainly built to support the notion of data and metadata. Nowadays the notion of service or “as a service” is increasing in the EO sector. The current granular questions around the GEO and FAIR principles do not address as such this notion of service or “as a service”. There is for sure room for improvement and e-shape could contribute to this task thanks to the large sample of individual DMPs that have been created by a representative panel of stakeholders in the 7 thematic showcases.
- The e-shape project has developed a tool, (An Excel spreadsheet including macros) that has proven to be an easy and efficient mean to launch the DMP information collection process at the pilot level. Nevertheless to allow a wider usage in the EO community a change of paradigm would be suitable. This is explained in more detail in section 4.2 of this report.

Capacity building resources on DMP

- DMP self-assessment tool: <https://gkhub.earthobservations.org/records/rtdy9-qnd28>
- GEO Dialog series on the Data Management Principles: <https://gkhub.earthobservations.org/search?q=dialogue&l=list&p=1&s=10&sort=bestmatch>
- Dialogue on the Data Management Self-Assessment tool : <https://gkhub.earthobservations.org/records/kgbd8-58w14>:
- Webinar on GEO Data Management and Sharing Principles 06 déc. 2019 [Data Management and sharing principles](#)

References on DMP

- GEO Data Management Principles: Data Management Principles Implementation Guidelines, GEO-XII – 11-12 November 2015. https://www.earthobservations.org/documents/geo_xii/GEO-XII_10_Data%20Management%20Principles%20Implementation%20Guidelines.pdf
- FAIR Data Management Principles: Turning FAIR into a reality, European Commission, Final Report and Action Plan from the European Commission Expert Group on FAIR Data, 2018. https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf
- e-shape D1.6 Data Management Plan public deliverable (ADD PUBLIC LINK)
- DMP-generation platform <https://dmponline.dcc.ac.uk>

Box 13-6: Data Management Plan. Lessons learnt from Showcase 4: [MyEcosystem](#) Pilot 2: [mySITE](#) (data provision, visualisation tools, and ecosystem status indicators)

The [mySITE](#) pilot contributed to the development of an overall Data Management Plan for the pilot addressing the Geo Data Management Principles as well as the FAIR Principles. This was reviewed by the e-shape project and provides the basis for the operation and further extension of the service developed.

Lessons learned:

- addressing the principles for FAIR and open data management is a good means for the further development of the services provided as well as to ensure communication with the stakeholders.

13.8 Quotation and Billing

Reliable billing of course is critical. One e-shape pilot has met some problems with undue high bills from a DIAS due to a process failure during the weekend. It could be solved with the DIAS support team but the experience has been annoying enough for the pilot to develop a strategy to avoid meeting the problem again.

In the context of a distributed architecture involving diverse cloud computing resources with their own specific quotation and billing system and strategy, estimating and optimizing the costs on the user side and implementing a global billing, eventually with different currencies on the provider's side require a unified billing solution. The OGC testbed 14 has addressed this issue to satisfy ESA needs. It looks reasonable to think that this will be more needed in the future to take advantage of public and private resources seamlessly.

References on Billing:

- OGC Testbed-14: Authorisation, Authentication, & Billing Engineering Report OGC 18-057 : <http://www.opengis.net/doc/PER/t14-D010>

13.9 The value of Standards, Data Models and Best Practices

13.9.1 What is the value of open standards?

Open standards are technical standards developed through a collaborative, consensus-driven process and are publicly available for anyone to use and implement. These standards are essential for promoting interoperability and ensuring that different systems can work together seamlessly.

The value of open standards lies in their ability to foster innovation, competition, and efficiency in various domains. Open standards promote healthy and fair competition and help to prevent the emergence of monopolies that stifle innovation.

Open standards also make it easier for companies and individuals to switch between different products and services, which lowers the risk of dependency, helps to reduce the costs of switching, and increases consumer choice. This, in turn, encourages companies to focus on improving their products and services to stay competitive, rather than relying on vendor lock-ins to protect their market position.

In addition to their economic benefits, open standards can also help to promote social values such as accessibility, transparency, and inclusivity. By making technical specifications publicly available, open standards ensure that all stakeholders can participate in the development process and that the resulting products and services are accessible to all users, regardless of their location, language, ability, fostering a real diversity including from underrepresented groups.

Overall, open standards play a critical role in promoting innovation, competition, accessibility, cost savings, social values, and risks resilience and their importance will only continue to grow as technology continues to advance and become more pervasive in our daily lives.

Some challenges to developing standards for new technologies are implementing a consensus-driven agile development process involving a wide diversity of stakeholders, adapting rapidly to a changing environment, maintaining some stability to allow time for broad adoption still evolving as new data sources become available and technology and needs evolve.

13.9.2 Open Standards and Earth Observations

Earth Observation (EO) is a field that uses various technologies along the value chain from data production, collection, management, processing and delivery. To ensure interoperability and compatibility between different EO systems, data products, and technologies along this value chain, a range of open standards have been developed and adopted by the EO community. Some of the key

standards used in EO include standards from the International Standard Organization - ISO-, the World Meteorological Organization - WMO-, and the Open Geospatial Consortium - OGC-. In Europe, the INSPIRE Directive, establishing an infrastructure for spatial information in Europe to support Community environmental policies, and policies or activities which may have an impact on the environment entered into force in May 2007 and have pushed the implementation of standards to make environmental data more accessible.

Annex 5 reviews the open standards implemented by the e-shape pilots.

Standards Architecture Diagram

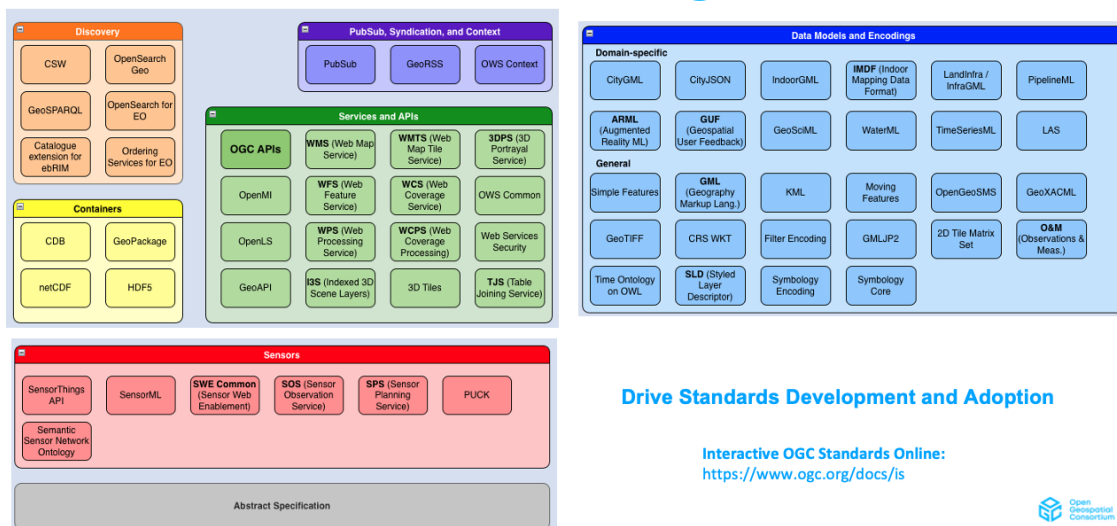


Figure 36: OGC Standard Architecture Diagram

Box 13-7: Standards, Data Models and Best Practices are useful for the success of each FAIR Principles: Findability, Accessibility, interoperability, and Reusability. Lessons learnt from mySITE

The [MyEcosystem](#) [mySITE](#) pilot has contributed to the improve the usability, accuracy, and currentness of data. e-shape has supported a process that connected real users to application providers. for [mySITE](#), the requirements involved "the implementation of standardised services for the provision of geodata and metadata to external services and users who incorporate site data in their systems and analysis workflows." The paper Designing and implementing a data model for describing environmental monitoring and research sites

13.10 Standards Compliance

13.10.1 Introduction

The quality of the standards implementation is key to reach real interoperability. Despite all the efforts made by the editors to avoid ambiguities and the concept of Reference Implementations to validate the usability and quality of the standards, the standards specifications can have some ambiguities leading to implementation variations that can impact the quality of interoperability. This is why it can be useful to use Compliance validation tools such as OGC Cite, the INSPIRE Validator, or NASA Compliance tool.

- e-shape developed a webinar on Compliance tools accessible on Youtube: Full webinar replay: Standards compliance Tools, benefits and return on experience <https://www.youtube.com/watch?v=BZX0O-pLxXE>
 - Extract on INSPIRE Validator : https://youtu.be/HD9_RfdaGRA
 - Extract on OGC Compliance Tool https://youtu.be/j3WCOZzfJ_U
 - Extract on the Nasa validator in use : <https://www.youtube.com/watch?v=LdVAtCJILCo>
 - References :
 - INSPIRE Reference Validator – Home : <https://inspire.ec.europa.eu/validator/>
 - OGC Compliance : <https://www.ogc.org/resources/compliance/>

13.11 Coherence and compliance of e-shape GEO DMP regarding the INSPIRE directive

As per the impact number five (I-5) and the corresponding assessment indicators, the e-shape project aims at addressing a "Coherent data management, through the use of GEOSS Data Management Principles and best practices (INSPIRE-compliant)". The means to monitor this impact is made via indicators to assess the "Percentage of INSPIRE compliant pilots within the project" and the "Percentage of pilots in full compliance to GEO Recommendations on interoperability and GEO Data sharing and Data management principles".

At pilots level the minimum requirement is to engage in GEO DMP discoverability (DMP-1) where "Data and all associated metadata will be discoverable through catalogues and search engines, and data access and use conditions, including licenses, will be clearly indicated" in association with INSPIRE which defines metadata recommendations and profiles.

In order to provide within e-shape an uniform and homogeneous approach, the project has benefited from the support of the webservice-energy.org GEO catalogue based on the open source GeoNetwork solution to create and host for each of the 37 pilots a unique metadata record based on interoperable and standards profiles (ISO 19139 and INSPIRE Network Services).

35 out of 37 metadata records have been created so far and are deployed in the webservice-energy catalogue: <https://tinyurl.com/5dk34cks>

Most of them are based on ISO 19139 metadata for dataset and service and one is using the INSPIRE Network Service implementation:

<https://inspire.ec.europa.eu/id/document/tg/metadata-iso19139/2.0/examples>.

Indeed to be valid the INSPIRE Network Service profile requires a "Contains Operation / Connect Point" information to be filled. This information is expected to link to an OGC: GetCapabilities operation. Among the e-shape pilots only S3P2 pilot "High photovoltaic penetration at urban scale" is offering direct access to such web service standard namely WPS (Web Processing Service) (See: <https://tinyurl.com/mwac4c5w>)

The webservice-energy catalogue is part of the GEO DAB (Discovery and Access Broker) targets list since 2008 and it is harvested on a weekly basis. Consequently the metadata records that are available and visible on the webservice-energy catalogue are visible as well on the GEO Web Portal search for "e-shape" or click directly here: <https://bit.ly/3MbO4Xw>

This workflow (standard metadata on standard catalogue -> harvest from DAB -> Search and discovery from the GWP) is fully addressing the GEO DMP-1 expectation.

Moreover in the scope of the project e-shape has worked in support of the GEO Knowledge Hub development team to fine tune GKH harvesting feature over GeoNetwork catalogues. This has led to the automatic harvest of the e-shape pilots metadata record available from the GKH. A dedicated

Community section has been created on the GKH to host all e-shape contribution (<https://gkhub.earthobservations.org/communities/e-shape>).

Regarding the INSPIRE compliance, the catalogue is offering 2 profiles, ISO 19139 and INSPIRE Network Service. The second is fully INSPIRE compliant and it passes the INSPIRE Validator Test as provide by the European Commission support application (<https://inspire.ec.europa.eu/validator/test-selection/index.html>).

INSPIRE Validator	
<XML> Metadata record following INSPIRE profile generated from the webservice-energy catalogue	4be5ce1e-5354-4195-9bcc-9c057e62399a.xml
Metadata record validation: Step #1	



Figure 37 : Metadata record validation: Step #1 screenshot

INSPIRE Validator

Metadata record validation: Step #2

The screenshot shows the INSPIRE Validator - Test reports page. The test run was on 18-06-2024 with test suite Conformance Class 4: INSPIRE Network Services metadata. The status is 'Passed, manual checks required'. The test run started on 20/04/2022 at 16:06:57 GMT and lasted 10 seconds. The test results table shows 3 test suites, 8 test cases, and 29 assertions, all of which passed. The page also lists the test suites: Common Requirements for ISO/TC 19139:2007 based INSPIRE metadata records, Conformance Class 3: INSPIRE Spatial Data Service baseline metadata, and Conformance Class 4: INSPIRE Network Services metadata. The report was generated by ETF.

Status	Passed, manual checks required
Started	20/04/2022 16:06:57 GMT
Duration	10 s

	Total	Count	Skipped	Failed	Warnings	Manual
Test suites	3	0	0	0	0	1
Test cases	8	0	0	0	0	1
Assertions	29	0	0	0	0	1

Report generated by ETF

Figure 38 : Metadata record validation: Step #2

Metadata record validation: Step #3

The screenshot shows the INSPIRE Validator - Test reports page with detailed test results for Conformance Class 3: INSPIRE Spatial Data Service baseline metadata. The test run was on 18-06-2024 with test suite Conformance Class 4: INSPIRE Network Services metadata. The status is 'Passed, manual checks required'. The test run started on 20/04/2022 at 16:06:57 GMT and lasted 10 seconds. The test results table shows 3 test suites, 8 test cases, and 29 assertions, all of which passed. The page also lists the test suites: Common Requirements for ISO/TC 19139:2007 based INSPIRE metadata records, Conformance Class 3: INSPIRE Spatial Data Service baseline metadata, and Conformance Class 4: INSPIRE Network Services metadata. The report was generated by ETF.

Status	Passed, manual checks required
Started	20/04/2022 16:06:57 GMT
Duration	10 s

	Total	Count	Skipped	Failed	Warnings	Manual
Test suites	3	0	0	0	0	1
Test cases	8	0	0	0	0	1
Assertions	29	0	0	0	0	1

Report generated by ETF

Figure 39 : Metadata record validation: Step #3

The ISO 19139 profile as available as a default profile from the webservice-energy catalogue implementation (GeoNetwork software), is not fully passing the INSPIRE Validator Test though most of the validation tests are compliant. In practice those two metadata standard profiles are complementing each other.

While the more formal INSPIRE one is expecting mandatory fields like web services GetCapability information to be part of the record file and consequently pass the Validator Test, the ISO 19139 offers more flexibility and provide less constraints regarding associated information to be part of the metadata record files, while on the contrary not passing the Validator Test.

In the framework of e-shape it is important to provide a certain amount of flexibility for addressing the collection of information for each of the 37 pilots in the 7 showcases. Not all pilots has for example provided a standard invocable web services in support to their application. Therefore the ISO 19139 will perfectly suited for this approach. Unlike, pilots that will provide OGC standard web service support to their pilot applications will benefit from the more formal and validated INSPIRE Network Service profile.

It is important to mention that whatever the preferred metadata profile for each pilot, the above-mentioned workflow will enable a fully interoperable approach empowering a wider search and discovery of e-shape assets for the benefit of the EuroGEO and GEO community.

This task has been implemented to address all the 37 pilots and to make sure that each of them benefit from a standard and interoperable metadata record as a support from the e-shape project.

13.12 Pilot Exploitation Readiness Level - PERL

The e-shape project brings forward the novel Pilot Exploitation Readiness Level (PERL) approach that builds on D5.2 “First PERL definitions” and the introduction of the TRL (Technology Readiness Level) in assessing the readiness of the e-shape pilots’ services.

Focused on technology, the TRL approach, is well accepted and used and has been customised to fit specific needs, but it doesn’t take into account the market readiness of a given solution. The methods to assess “readiness” have shifted from technology readiness to go-to-market readiness, and methodologies increasingly consider both technology, market, and commercialisation approaches. Different actors are considered to evaluate the maturity of the solutions; however, the market-oriented perspective could generate difficulty for a non-commercial user such as Research and Development - R&D sector involved in H2020 projects.

Clear guidelines with a clear definition of criteria for a dedicated solution are needed to ensure coherent assessments. In this context, the novel Pilot Exploitation Readiness Level (PERL) approach that will be described in detail in this deliverable, is to establish a standardised methodology for the assessment of the maturity of R&D activities and the potential of the company/institute to valorise, exploit and successfully deliver these solutions. The PERL methodology has a double value: it proves to be a tester methodology with 32 Pilots projects as candidates; and an open door for the results that can be used as a standard for those entities that need to evaluate their service market readiness.

The Pilot Exploitation Readiness Levels (PERL) methodology aims to support the development of R&D activities and in particular to support the identification of milestones to be reached for a solution to become sustainable and, or to reach the market. In that regard, the PERL would be used as a metric to:

- Assess the current status of e-shape Pilots;
- Assess potential new Pilots to be on-boarded;
- Identify technology or business components to be developed or improved;
- Apply to other R&D projects to promote the exploitation of their results and their transformation into market-ready products.

The PERL is considered to be a ‘living’ metric, ready to adapt to the rapidly changing EO services marketplace. As such, the PERL has undergone substantive changes since the publication of D5.2, resulting in a leaner approach, a more efficient data-gathering process, and more easily comprehensible results. The PERL concept continues to evolve both within and beyond the course of the e-shape project

as we attempt to categorise and assess Pilot maturity both for the purposes the e-shape project and more generally across the sector. The parameters of the updated PERL are broken down into three main categories (“TOM”):

- Technology: The technical ability to deliver the solution and keep it up to date.
- Operations: The internal team skills & processes needed to bring the technology to the identified market.
- Market: The readiness of the team to capture the relevant markets for the solution (knowledge of customer needs, pricing strategy, business plan, competition etc.)

The raison d’être of “TOM” is to provide a detailed overview of the service readiness from the technological maturity of a given pilot through the assessment of a supply chain establishing the steps of the service’s uptake towards its market exploitation and sustainability. These three categories fit both with the range of service readiness reviewed and with the key support mechanisms offered through the e-shape project to Pilots.

The healthy approach to the development of a new index, in which the PERL remains a ‘live’ concept, ready to adapt to a rapidly changing marketplace, has the potential to become a methodology of international validity and will be one of the keys to its success in being taken up by the EO and other sectors to describe the readiness of a new service to enter the market.

References on PERL- Pilot Exploitation Readiness Level: <https://e-shape.eu/index.php/resources#sppb-modal-1679641282100>

14 ANNEXES SHORT DESCRIPTION

14.1 Annex 1: e-shape pilot applications

This annex gathers the description of the showcases and the links to the pilots' descriptions on the e-shape Website for the convenience of the reading.

Link: <https://shorturl.at/jyMTW>

14.2 Annex 2: Glossary

The glossary provides the meaning of the acronyms used in the Best Practices.

Link: <https://shorturl.at/muRVZ>

14.3 Annex 3: Copernicus Services used by the e-shape pilots

The Objective O2 of the project mentions “it is critical to exploit the IT capabilities and the wealth of data made available through DIAS, GEOSS platform, NextGEOSS, EOSC, in-situ observatories (as organized in ENVRI plus), citizen observatories, and any other existing hubs or platforms. » Additionally, the project KPIs included “Number of Copernicus datasets exploited in the Showcases pilots. » and in the Initial assessment, some partners asked to provide some Data assets analysis. To address all these issues, this annex captures the Copernicus, in situ and other major datasets used by the pilots (excluding here some datasets more specific to each pilot). When relevant, we have listed the datasets used per the pilots and reverse the pilots using each dataset. This shows for instance that the Digital Elevation Models provided by CLMS, ERA5 provided by C3S are very used. It also shows that CMEMS data are not only used by the Water Resources Management pilots but also by some Renewable Energy, Biodiversity, Disaster Resilience and Climate pilots proving the broad value of these data.

Link: <https://shorturl.at/bjDW3>

14.4 Annex 4: Open source software and packages used by the e-shape pilots

This Annex collects the Open Software and other software most frequently used by the pilots and a short description of their use. The number of pilots using them can be considered as a criteria of their usefulness and maturity.

Link: <https://shorturl.at/HKSTV>

14.5 Annex 5: Standards used by the e-shape pilots

This Annex identifies the Standards most used by the pilots identifying a baseline of the most useful standards. The OGC APIs have been published more recently.

Link: <https://shorturl.at/glpqx>

14.6 Annex 6: Platforms used by the e-shape pilots

This Annex identifies the platforms used by the pilots and, when known, what they are used for : Data access, Processing, Publication, outreach.

Link: <https://shorturl.at/vJKY5>