Project Title	Towards an Interoperable MARinE Knowledge GRAPH
Project Acronym	MareGraph
Grant Agreement No.	101100771
Start Date of Project	23/01/2022
Duration of Project	36 Months

4.1 - Architecture Whitepaper

Work Package	WP4 – Project Management and Coordination
Lead Author (Org)	Julián Rojas (IMEC)
Contributing Author(s) (Org)	Marc Portier (VLIZ), Bart Vanhoorne (VLIZ), Vidyashree Tarikere (IMEC)
Due Date	M6
Date	30.06.2023
Version	V1.0

Funded by the European Union

Dissemination Level

☒ PU: Public

□ SEN - Sensitive

□ RE: Restricted to a group specified by the consortium (including the Commission)

□ CO: Confidential, only for members of the consortium (including the Commission)

Versioning and contribution history

Version	Date	Author	Notes
0.1	16.05.2023	Julián Rojas (IMEC)	Initial draft for partner review
1.0	17.07.2023	Julián Rojas (IMEC)	First complete version

Disclaimer

This document contains information which is proprietary to the MareGraph Consortium. Neither this document nor the information contained herein shall be used, duplicated, or communicated by any means to a third party, in whole or parts, except with the prior consent of the MareGraph Consortium.

Funded by
the European Union

Funded by the European Union

Table of Contents

TERMINOLOGY

Terminology/Acronym	Description
LDES	Linked Data Event Stream
LDF	Linked Data Fragment

Funded by the European Union

TPF	Triple Pattern Fragment
API	Application Programming Interface
MR	Marine Regions
WoRMS	World Register of Marine Regions
EuroBIS	European Ocean Biodiversity Information System

Executive Summary

This white paper provides a conceptual and technical background around the motivations for the conception of the MareGraph project. It also describes the rationale for the technical design choices made to support the proposed reference architecture that is detailed here. Finally, hints towards some open technical challenges.

1 Introduction

In 2019 the EU parliament published the directive 2019/1024[1], also known as the Open Data Directive (ODD). It served as an amendment of the previous directive 2003/98/EC[2] on the reuse of public sector information, and it went on to further recognize the important role of open data in promoting social engagement, and kick-start and promote the development of new services based on novel ways to combine and make use of such information. In this directive, open data is understood as any data in an open format that can be freely used, re-used and shared by anyone for any purpose, in alignment with the definition given by the open data charter[3]. The main call was to encourage Member States to promote the creation of data based on the principle of *open by design and by default*, meaning that data should be widely available for re-use both in the public sector and for commercial purposes with minimal or no legal, technical or financial constraints.

An important type of data, recognized to fall under this definition, is research data. The volume of research data generated keeps on growing at an exponential rate and has potential for re-use within research and beyond the scientific community [1]. Research data includes statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L1024
[2] https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A32003L0098
[3] https://opendatacharter.net/

and images. It also includes meta-data, specifications and other digital objects that should be made accessible as early as possible in the dissemination process to facilitate its use and re-use. The principal goals are helping to enhance quality, reduce the need for unnecessary duplication of research, speed up scientific progress, combat scientific fraud, and it can overall favor economic growth and innovation. The adoption of related efforts, such as the FAIR (Findable Accessible Interoperable Reusable) principles[4] are also encouraged and sometimes required. ￼

A number of thematic categories of high-value datasets were listed by the ODD, together with a set of requirements for their publishing as seen in Figure 1. From a technical perspective, (dynamic) data should be made accessible through well-designed APIs and/or bulk downloads. The set-up and use of APIs need to be based on several principles: availability, stability, maintenance over lifecycle, uniformity of use and standards, user-friendliness as well as security. Public sector bodies and public undertakings should promptly provide dynamic data, which refers to frequently updated data, sometimes in real-time. This information should be made available for re-use soon after its collection, utilizing appropriate APIs and, if applicable, as a bulk download. The only exception to this rule would be cases where publishing the data would require excessive effort in proportion to the expected benefits. This hints towards the inherent complexity and associated costs that setting APIs might bring to a public sector

[4] https://zenodo.org/record/3909563#.YoIj4mBBzKp

body, while also showing the broad definition attributed to the API concept.

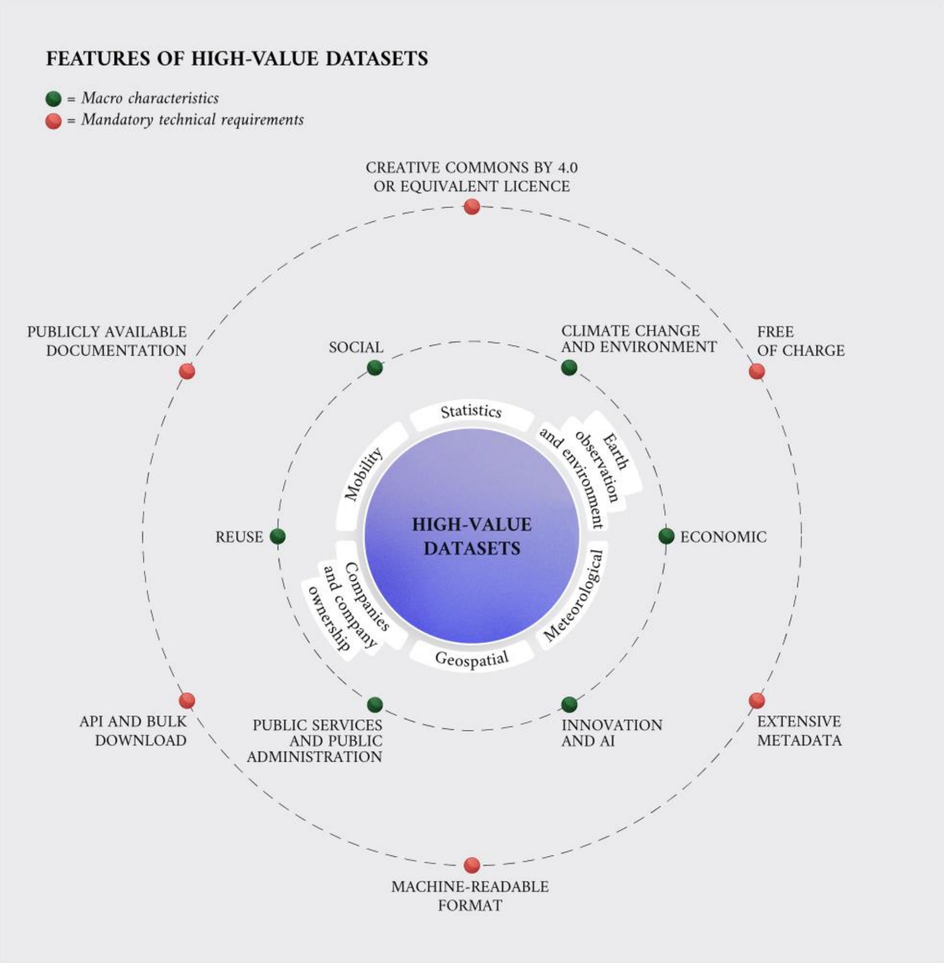

FEATURES OF HIGH-VALUE DATASETS
● = Macro characteristics
● = Mandatory technical requirements
CREATIVE COMMONS BY 4.0 OR EQUIVALENT LICENCE
PUBLICLY AVAILABLE DOCUMENTATION
FREE OF CHARGE
SOCIAL
CLIMATE CHANGE AND ENVIRONMENT
Statistics
Earth observation and environment
Mobility
REUSE
HIGH-VALUE DATASETS
ECONOMIC
Meteorological
Companies and company ownership
Geospatial
API AND BULK DOWNLOAD
PUBLIC SERVICES AND PUBLIC ADMINISTRATION
INNOVATION AND AI
EXTENSIVE METADATA
MACHINE-READABLE FORMAT

Fig 1. Publishing requirements and thematic categories of high-value datasets (source online[5]).

The requirements of the ODD show an overlap with the principles defined by the EU Interoperability Framework (EIF), published by the EU Commision in 2017[6]. We can observe direct similarities related to the openness and reusability principles, and the advocacy for the application of standards to ensure seamless data flows and interoperability at different levels (Figure 2). A *Legal Interoperability* level is mentioned, where coherence and removal of legal barriers is advocated through "interoperability checks" of legal frameworks within and across Member States. The *Organizational Interoperability* level prompts for the alignment of business processes that clearly communicate the responsibilities and expectations regarding digital information services of public bodies. The *Semantic Interoperability* level

[5] https://data.europa.eu/en/publications/datastories/high-value-datasets-overview-through-visualisation
[6] https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf

Funded by the European Union

urges to ensure precise and explicit meaning of data and information exchanged among organizations. A direct reference to Linked Data technologies and the employment of standard and well-known vocabularies/data models is given as a recommended strategy for information management. Lastly, the *Technical Interoperability* level refers to the use of common protocols and communication interfaces that facilitate systems interlinking both within and across organizational boundaries. The adoption of formal and open technical specifications is advised. It is at this level where an initial call for the employment of reusable and well-defined APIs can be observed, which is later reiterated and made more explicit in the ODD.

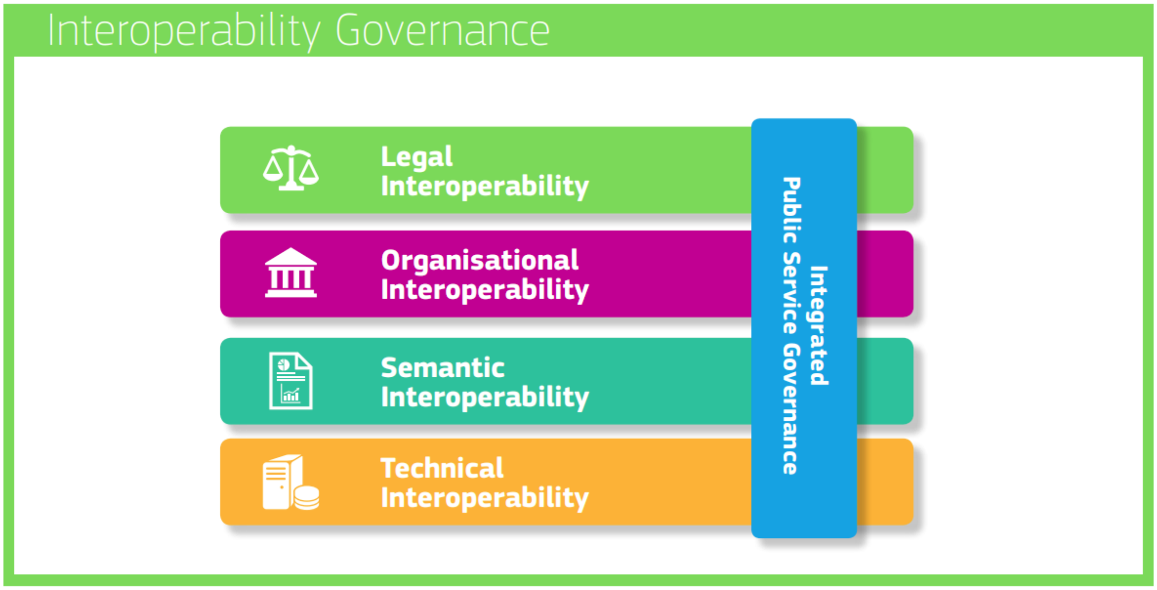

Interoperability Governance
Legal Interoperability
Organisational Interoperability
Semantic Interoperability
Technical Interoperability
Integrated Public Service Governance

Fig 2. Layered interoperability model of the EU Interoperability Framework (figure taken from the EIF).

Further steps were taken on the European Strategy for Data[7](ESD) published by the EU Commision in 2020, and later in the EU Data Act[8] (EDA) from 2022. The ESD introduced the vision for a single European data market, guided by principles of improved access and responsible usage of data (Figure 3). The term *Data Space* appears first in this document as a reference to such single market for data, where personal as well as non-personal data, including sensitive business data, are secure and businesses also have easy access to a large amount of high-quality industrial data, boosting growth and creating value. This Data Space vision has as a main objective to address multiple challenges including:

- *Data accessibility and availability*: Data is often concentrated in the hands of a few dominant players, limiting access and hindering competition and innovation. It emphasizes the need to

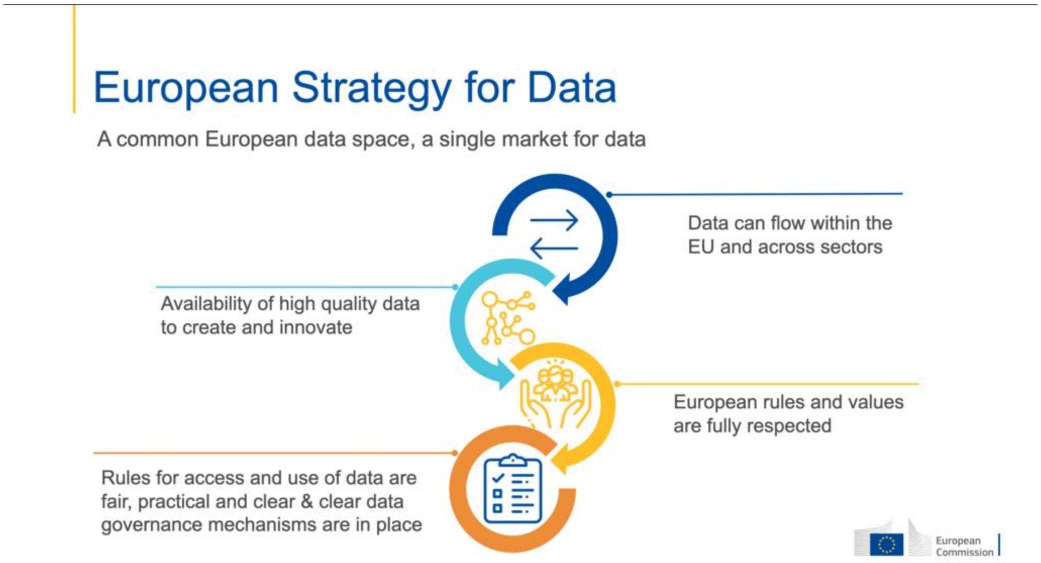
[7] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1593073685620&uri=CELEX%3A52020DC0066
[8] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN

Funded by the European Union

ensure fair and open access to data, particularly for small businesses and startups.

- *Data interoperability and standards*: The lack of harmonized standards and interoperability across different data sources and systems. It highlights the importance of developing common standards and promoting interoperability to facilitate seamless data sharing and integration.
- *Data governance and trust*: The strategy acknowledges that the lack of trust in data practices can hinder data sharing and collaboration. It emphasizes the need to establish clear rules and frameworks for data governance, including data protection, privacy, and security measures, to build trust among individuals, businesses, and public administrations.
- *International cooperation and data flows*: The strategy recognizes the global nature of data and the need for international cooperation on data governance. It acknowledges the challenges related to cross-border data flows, data localization requirements, and differing legal frameworks across countries, and calls for the establishment of international data partnerships and common standards.

The EDA elaborates on this vision and brings forward a regulatory framework that sets harmonized rules on how data can be accessed and used and provides safeguards for business and individuals to protect their data. Among the key provisions of the EDA is the requirement for business to provide users with more control, including the right to access, port and delete their data. The development of common sector-oriented Data Spaces is also specified through a number of requirements that include (i) transparency and openness for businesses and organizations in general to share data on fair, reasonable, and non-discriminatory terms; (ii) clear rules to govern and protect the privacy and security of data; and (iii) interoperability that allows to easily share data across different data spaces. Accomplishing this vision would ensure that more data becomes available for use in the economy and society, while keeping companies and individuals who generate the data in control.



European Strategy for Data

A common European data space, a single market for data

Data can flow within the EU and across sectors

Availability of high quality data to create and innovate

European rules and values are fully respected

Rules for access and use of data are fair, practical and clear & clear data governance mechanisms are in place

European Commission

Fig 3. Goals of the European Strategy for Data towards a common EU data space (online source[9]).

This observed evolution on the understanding and conceptualization of data management in the European context, set the conditions and provided the motivation for the MareGraph project. This project aims to abide by the recommendations brought forth by the ODD and the EIF, while following the guidelines of the ESD and the EDA, to increase the Legal, Technical and Semantic interoperability of three particular datasets related (but not limited) to the marine biology domain:

- a taxonomic dataset (World Register of Marine Species - WoRMS)
- a geospatial dataset (Marine Regions)
- a biogeographic dataset (EurOBIS)

These datasets fit under the *Geospatial* and *Earth Observation and Environment* categories of high-value datasets defined by the ODD (Figure 1), which highlights their importance and the need for well-defined lifecycle management and governance strategies. This is precisely the central element of MareGraph, where one of its main goals is to **design and implement a Linked Open Data-based strategy for the production and publication** of the WoRMS, Marine Regions and EurOBIS datasets. Extending the historical role of these datasets in the field of semantic interoperability is essential to ensure their continued dissemination and applicability in the marine domain and the application of Linked Open Data

[9] https://dataspaces.info/common-european-data-spaces

principles has the potential to effectively enable their conceptual and practical integration, thus increasing the possibilities of reuse and making a significant contribution to the creation of a **European Marine Knowledge Graph (KG)**.

The MareGraph project tackles the challenge of increasing interoperability on different levels, as defined by the EIF (Figure 2). At a *Legal* level, the task is set for defining clear license and reuse policies in accordance with the existing Data Space protocols, that for example, can handle the complexity of certain subsets of data not being entirely open and are restricted for commercial purposes, as is the case of EurOBIS. At an *Organizational* level, collaboration between cross-national organizations foresees the definition of unified processes that combine and extend existing national methodologies to address the creation and governance of a European Marine KG. This includes the definition of formal ontologies, following best practices based on reuse of existing standard vocabularies, thus addressing the *Semantic* interoperability level. Finally, at the *Technical* level, the focus is on the definition of an architecture that fulfills the following requirements:

- *Cost-efficient*: A data publisher needs to be able to publish data on a small budget, and still make sure the potential in the ecosystem remains the same.
- *Flexible*: Developers need to be able to answer any query and speed up certain kinds of queries on their own, while the data publisher should be able to evolve their APIs independently.
- *Sustainable*: The datasets need to be able to be kept online as part of the core tasks of the organizations managing these datasets.

In this white paper, we focus on detailing the design of such architecture. The main goal is therefore to provide a frame of reference for implementing highly available and interoperable dataset production and publication systems, able to maximize the number of interaction interfaces with the data (APIs), while keeping associated costs at reasonable levels.

In the remainder of this paper, we first present a description of the related concepts and technologies used in the definition of our reference architecture: Linked Open Data (chapter 2), Linked Data Event Streams (chapter 3) and Data Spaces (chapter 4). Then we provide a detailed view on the design of the architecture (chapter 5) and we conclude with some final remarks (chapter 6).

2 Linked Open Data

Linked Open Data (LOD) encompasses all data that is openly published on the Web and that adheres to a set of best practices known as the Linked Data principles [2]. These principles aim to establish standards for the representation and accessibility of data on the Web. Additionally, they promote the creation of hyperlinks between data from diverse sources. By interconnecting these hyperlinks, Linked Data forms a unified global data graph, analogous to how hyperlinks on the traditional web connect HTML documents to create a comprehensive global information network [3]. The publication of LOD is guided by the following four principles:

1. **Use URIs as names of things**: A uniform resource identifier (URI) is a string of characters that serves as a unique identifier for various types of things, including digital content, tangible objects, or abstract concepts. URIs enable to differentiate between different entities while also recognizing those that are the same. For instance, an entity may have different names in different languages, but its URI remains constant. To ensure the continuity of its meaning, a URI needs to be persistent, meaning it is permanently associated with a specific resource.
2. **Use HTTP URIs so that people can look up those names**: LOD commits to use the HTTP protocol to allow data sources to be accessed using generic data applications such as browsers, search engines, etc.
3. **When someone looks up a URI, provide useful information, using the standards (RDF and SPARQL)**: LOD entails publishing machine-readable and easy to interlink data. It is then crucial to use a standard format to represent the data and to use a standard query (search) language. The W3C recommendations RDF and SPARQL fulfill this role as standard solutions.
4. **Include links to other URIs to discover more things**: URIs guarantee entity identifiers to remain globally unique. This enables hyperlinks to be set between entities in different data sources. These (RDF) links connect all LOD into a single global data graph and enable applications to discover new data sources on the fly.

Although the LOD principles mention the SPARQL standard as a recommended mechanism to access and query Linked Data on the Web, it is important to recognize the difference of the conception of SPARQL as a query language and as a data access interface. On the one hand, the SPARQL Query Language[10] defines the standard syntax and semantics by which RDF data can be uniformly queried, regardless of the physical location of the data (i.e., local or remote). On the other hand, the SPARQL Protocol[11] defines a means for conveying SPARQL queries and updates to a (remote) SPARQL processing service and returning the results via HTTP to the entity that requested them. This protocol assumes a traditional client-server setup where the server, in this case a SPARQL processing service, is responsible for processing and resolving every received query. In the next sections, we will describe alternative Linked (Open) Data publishing approaches that make different assumptions on how query execution responsibility is distributed between clients and servers on the Web and in turn, establish different

[10] http://www.w3.org/TR/sparql11-query/

[11] https://www.w3.org/TR/sparql11-protocol/

Funded by
the European Union

trade-offs on publishing costs, query flexibility and performance.

2.1 Publishing Linked Open Data on the Web

Before the Linked Data initiative [4], the Semantic Web suffered from a causality issue: there were no applications because there was no data, and there was no data because no applications were using it. However, since Tim Berners-Lee's call for "Raw data now", more knowledge graphs exist as Linked Data than ever before[12]. Thus, the ball is now back in the Semantic Web's court: given this increasing amount of data in various domains, it should be possible to build the envisioned intelligent applications [5].

The presence of live queryable knowledge graphs on the Web appears to remain limited. "Live queryable" refers to Linked Data that can be queried directly without the need to download the entire knowledge graph beforehand. By "low availability," we refer to the dual challenge currently faced by the Semantic Web: i) the majority of knowledge graphs are not made available in a queryable format [6], and ii) knowledge graphs that are accessible through public SPARQL endpoints (via the SPARQL Protocol) suffer from frequent downtime or are subject to query limitations [7]. This lack of availability becomes even more problematic when considering queries spanning multiple distributed knowledge graphs. Understandably, many publishers opt for the safer approach of offering data dumps instead of assuming the responsibility of hosting a SPARQL endpoint. However, this approach does not bring us closer to achieving the goals of the Semantic Web because such data dumps need to be downloaded and stored locally so that the actual querying can happen offline. Additionally, their utilization requires sufficiently powerful machines (thus excluding edge devices) and technical expertise to set up. Consequently, a significant portion of Linked Data knowledge graphs is therefore not reliably queryable, nor easily accessible, on the Web.

To materialize the vision of Semantic Web applications built on dynamic knowledge graphs, we need to reassess the approaches for publishing LOD at a Web scale. The traditional approaches of data dumps and SPARQL endpoints represent two extremes, but there exists an unexplored range of potential Web interfaces in between. Our focus is to explore solutions that strike a balance between minimal server complexity, reducing the burden on data publishers, and facilitating live querying capabilities,

[12] 1255 linked datasets are reported in the Linked Open Data Cloud as of 11/2022: https://lod-cloud.net/

Funded by the European Union

maximizing the usefulness for Semantic Web applications.

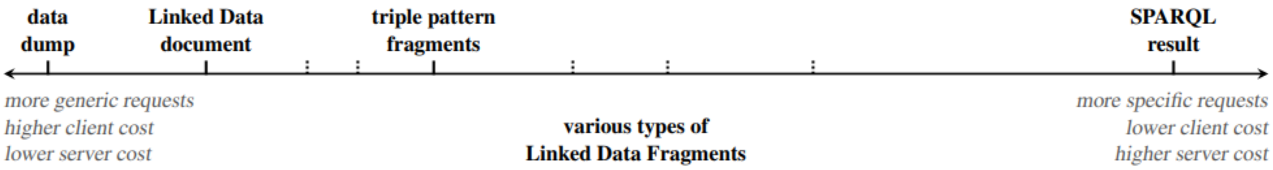


Fig 4. Every RDF Web API offers a set of Linked Data Fragments from a knowledge graph. These fragments differ in the specificity of the data they contain, and thus the cost to create them [8]**.**

Figure 4 portrays a uniform view on different Web APIs to publish and query LOD based on the granularity of selection mechanisms through which they expose an underlying knowledge graph. Such view is derived from observing what all such APIs have in common: in one way or another, they publish specific *fragments* of a knowledge graph. A SPARQL endpoint response, a Linked Data document, and a RDF data dump, each offer specific parts of all triples/quads of a given knowledge graph. Informally, a fragment of a knowledge graph is a resource consisting of a **specific subset of RDF triples/quads** of this graph, potentially **combined with metadata**, **and hypermedia controls** to retrieve related fragments. This conceptual description is known as the Linked Data Fragments (LDF) framework [8] and introduces the notion of possibly an infinite amount of LDF APIs with varying levels of granularity and thus different cost trade-offs for publishers and consumers, existing in between the more traditional data dumps and SPARQL endpoint APIs.

2.2 TREE Hypermedia Specification

Following the LDF conceptualization, the TREE hypermedia specification[13] was introduced. TREE conceives a dataset as a collection of data entities or *members* that may be fragmented in diverse ways, going beyond simple pagination [9]. Each collection fragment is published via HTTP together with metadata that includes explicit and qualified relations (in the form of HTTP URLs) to other related fragments. The overall view emerging from these relations could result into data structures similar to the organization of, among others, classic search trees, linked lists or skip lists, usable as an effective guide to navigate to the best fitting fragments that are relevant for answering a particular type of query. For example, by configuring a fragmentation of a dataset resembling a prefix-tree[14] based on text properties of its member entities, it becomes possible for a client application to autonomously traverse this data structure to find all entities with labels that start with a given prefix. Other types of fragmentations can be foreseen according to the nature and common use cases of datasets. The TREE

[13] https://w3id.org/tree/specification
[14] https://en.wikipedia.org/wiki/Trie

Funded by the European Union

specification defines an ontology and a set of types that provide the means to describe and configure such fragmentations. A concise overview of the specification is depicted in Figure 5.

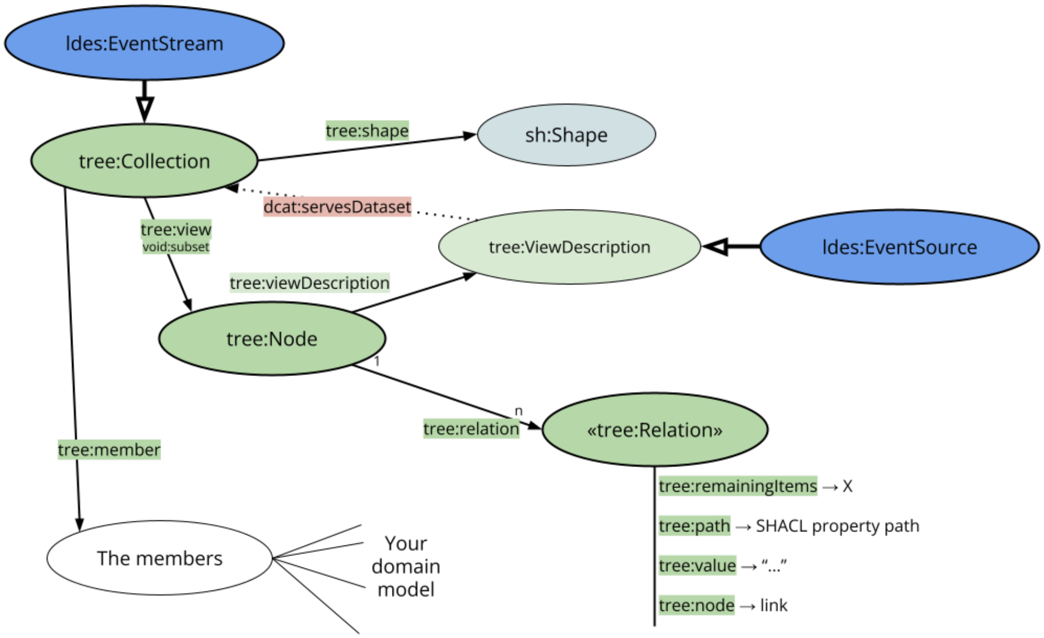


Fig 5. Overview of the TREE ontology. The `tree:Collection` class constitutes the central concept of the ontology and represents a given dataset that may be fragmented in various ways.

The core collection design is largely inspired by and in line with Hydra hypermedia vocabulary[15]. An instance of a `tree:Collection` links using the `tree:member` to the entities it contains. A fragment is defined by the `tree:Node` type. When all members of a collection can be found when starting from a `tree:Node`, the node may be considered as a *root fragment/node* of the collection, and it is linked to it with the `tree:view` property. Relations between nodes are defined by means of the `tree:relation` property and specified by entities of type `tree:Relation` or subclasses thereof.

Listing 1 shows an example of a TREE collection that provides one view identified by the `ex:ThisPage` node. This node publishes a data entity (`ex:entity1`) with its properties. It also defines a relation (`ex:Relation1`) towards another node, namely `ex:NextPage`, which publishes another data entity (`ex:entity2`). This particular relation includes a semantic description that expresses its nature. In this case it describes that all collection members present in `ex:NextPage` have a value greater than

[15] https://www.hydra-cg.com/spec/latest/core/

Funded by the European Union

5.1234 for the predicate `ex:property1`. The relation type `tree:GreaterThanRelation` is a subclass of `tree:Relation` which conveys this specific semantic interpretation. TREE defines various others concrete types of relations related to common use cases (e.g., substring, geospatial, etc.).

```
@prefix tree: <https://w3id.org/tree#>.
@prefix ex: <http://example.org/>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix sosa: <http://www.w3.org/ns/sosa/>.

####### http://example.org/myCollection ######
ex:myCollection a tree:Collection ;
      tree:view ex:ThisPage .

####### http://example.org/ThisPage ######
ex:myCollection tree:member ex:entity1.

ex:entity1 a ex:DataEntity;
      ex:property1 "4.053"^^xsd:double .

ex:ThisPage a tree:Node ;
      tree:relation ex:Relation1 .

ex:Relation1 a tree:GreaterThanRelation ;
      tree:node ex:NextPage ;
      tree:value "5.1234"^^xsd:double ;
      tree:path (ex:property1) .


####### http://example.org/NextPage #######
ex:myCollection tree:member ex:entity2.
      ex:entity2 a ex:DataEntity ;
      ex:property1 "6.567"^^xsd:double .

ex:NextPage a tree:Node .
```

Listing 1. Example of a TREE collection description.

Furthermore, a `tree:Collection` is defined as a subclass of a `dcat:Dataset`. The specialization being that it adheres to this collection design. A collection and its subclasses should define what their members look like using a SHACL Shape[16]. This SHACL shape may prove useful when selecting datasets from a DCAT catalog, also called *dataset discovery.*

[16] https://www.w3.org/TR/shacl/

3 Linked Data Event Streams

The strategies in which a dataset can be fragmented are virtually infinite, similar to the number of indexes one can make on top of a relational database. One method of thinking about a fragmentation strategy is to only fragment a dataset from the moment a dataset becomes too large to fit in one HTTP document. The way in which a second page will be added can then be decided based on the needs of the ecosystem. The priority could be to build the relation to the second page based on how the dataset is growing. That way, such fragmentation can be efficiently used for populating and keeping *in-sync* other Web APIs on top of that dataset.

This is precisely the main design goal of a Linked Data Event Stream (LDES) specification[17] [10]. LDES is meant to allow data publishers to focus on their core task, which is maintaining and publishing the data itself – while enabling third parties to replicate the entire dataset into whichever service they desire. In technical terms, an Event Stream is a one-dimensional fragmentation of an ever-growing collection of immutable objects. Each time the definition of an entity changes, this becomes a new object in the collection with its own URI. Because all objects on a certain fragment are guaranteed to be immutable, fragments that are full can be labeled as immutable – enabling efficient caching of the entire collection (Figure 6). Semantically speaking, a LDES (`ldes:EventStream`) is defined as a subclass of a `tree:Collection` (Figure 5), thus relying on the TREE vocabulary to describe how its fragmentations are configured.

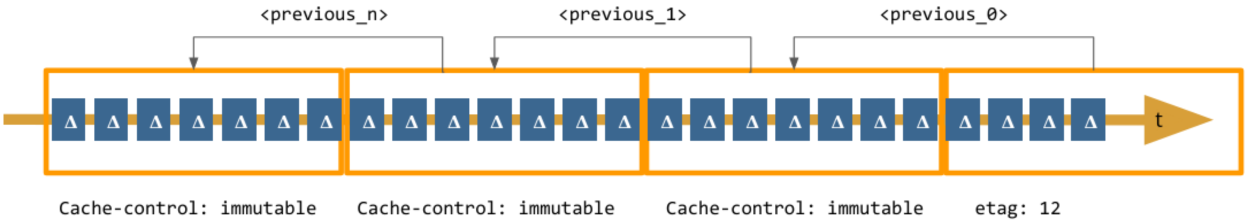


Fig 6. Visualization of a Linked Data Event Stream fragmented based on entity changes over time.

Conversely, data consumers can periodically poll the fragments that are not labeled as immutable to discover changes to a dataset relatively quickly, depending on how often they poll the data. LDES shows conceptual similarities with existing event streaming technologies, such as Apache Kafka[18] or RSS feeds[19]. Consumers can ingest an Event Stream to set up services the data publisher did not develop. Examples of such services include (Geo)SPARQL endpoints, and document stores such as ElasticSearch to provide full-text search over the data. Moreover, since every change is published as a separate object,

[17] https://w3id.org/ldes/specification
[18] https://kafka.apache.org/
[19] https://www.rssboard.org/rss-specification

Funded by the European Union

an Event Stream can be interpreted as a change log that can natively support publishing of historic data, often treated as an afterthought by data publishers.

From a data management perspective, LDES aims to become the core API for Linked (Open) Data publishing on the Web. LDES allows data publishers to focus on designing and maintaining only a minimalistic and relatively simple API, that in turn becomes an enabler for supporting further data workflows either within the publisher's premises or under the control of external third parties. In this way, an ecosystem of data exchange can be conceived (Figure 7), where data reuse is maximized and not constrained by the availability of resources or the willingness of data publishers to create and maintain particular types of APIs over their data.

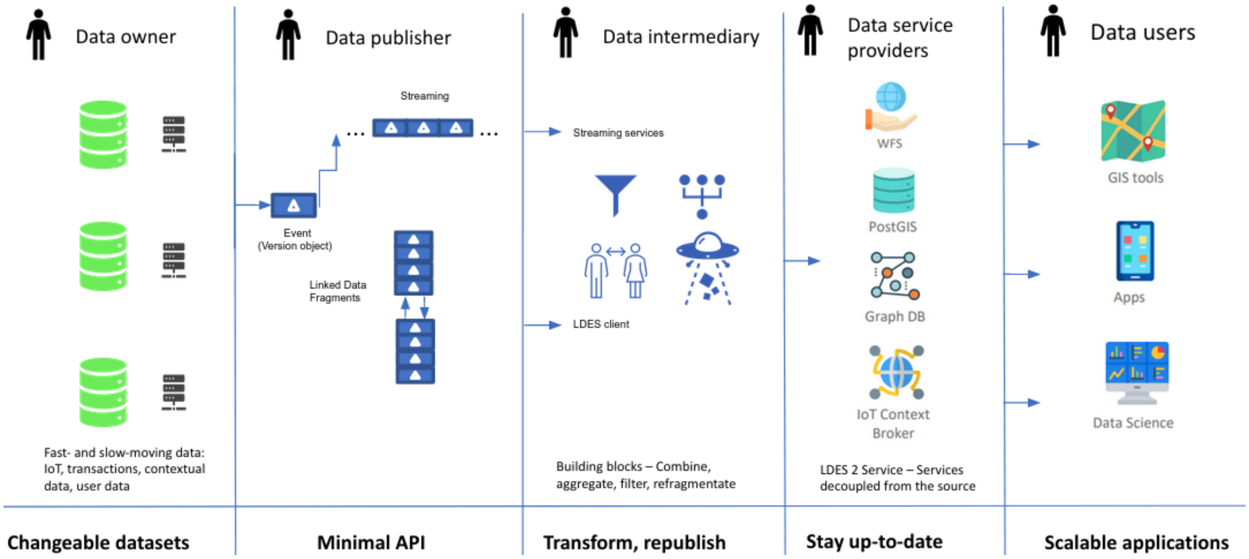


Fig 7. Role vision of a LDES-based ecosystem.

The LDES specification focuses on defining data publishing-related aspects that include, for example, the definition of retention policies[20]. These allow publishers to be transparent with their consumers and advertise how much and for how long an event stream will be kept online. Yet, data exchange among independent organizations entails an additional set of concerns (authentication, negotiation, usage control, etc.), as specified by the recent Data Spaces initiatives. In this context, LDES can be seen as a

[20] https://semiceu.github.io/LinkedDataEventStreams/#retention

Funded by
the European Union

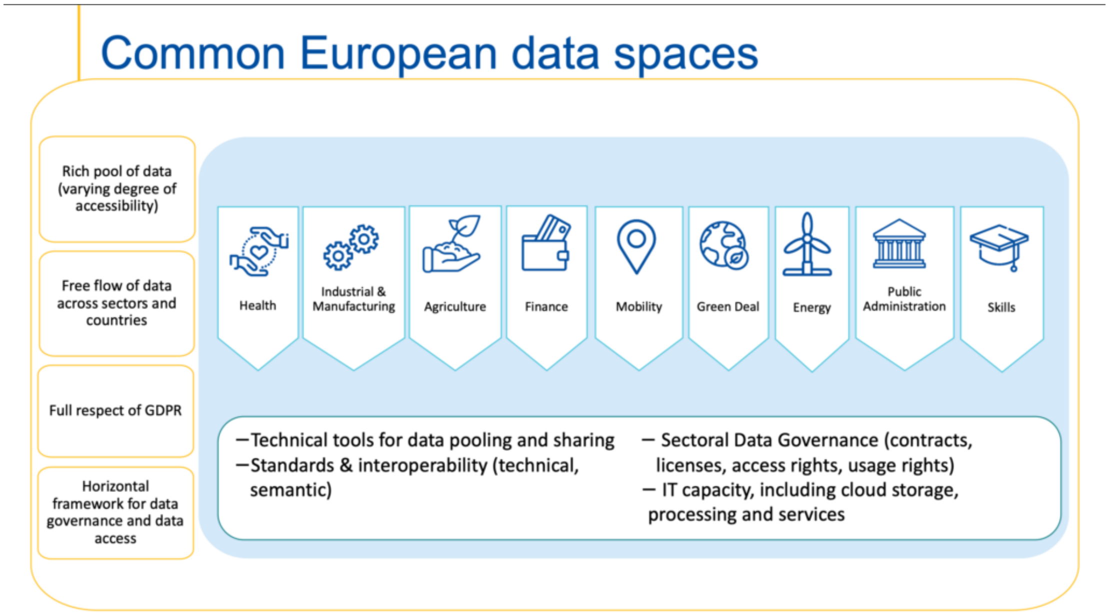
complementary building block centered on decentralized and sustainable data access.

Funded by
the European Union

4 Data Spaces

Data Spaces are the latest vision on how data transactions can take place in a way that is qualitative, reliable and secure, as well as scalable, economically viable, socially relevant and legally just. Our society and economy increasingly rely on the knowledge and revenue generated by generating, analyzing and sharing gigantic and diverse amounts of data – some more valuable or more sensitive than others.

Europe's goal is to be able to implement a better and more sustainable policy with the help of knowledge from data. The EU also wants to create a fairer playing field in the field of data-driven applications by democratizing access to data and creating a new balance in the associated costs and benefits. This should give smaller and emerging data companies the same opportunities as established players to tap into existing and new markets - and thus also allow more innovative applications and business models to emerge. As previously mentioned, this is materialized in the EU by the introduction of, among others, the ESD and EDA legal frameworks. These provide the definition of a set of initial common and sector-oriented Data Spaces, as seen in Figure 9.



Common European data spaces

Rich pool of data (varying degree of accessibility)

Free flow of data across sectors and countries

Full respect of GDPR

Horizontal framework for data governance and data access

– Technical tools for data pooling and sharing
– Standards & interoperability (technical, semantic)
– Sectoral Data Governance (contracts, licenses, access rights, usage rights)
– IT capacity, including cloud storage, processing and services

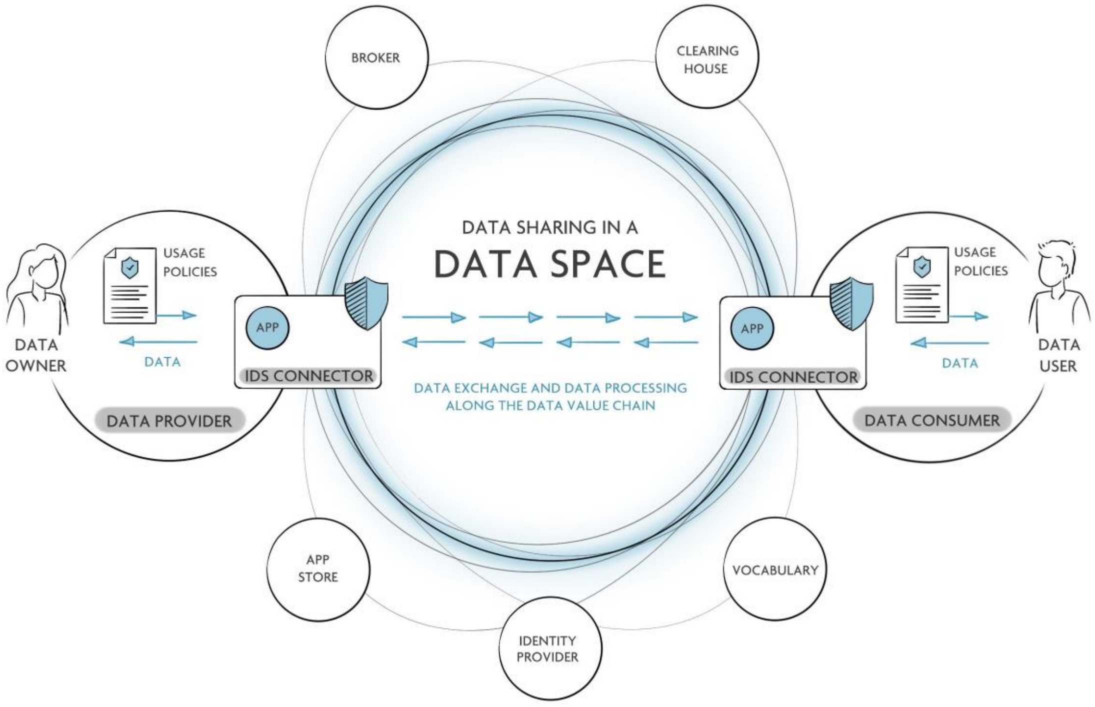
Fig 8. Initial definition of common sector-oriented European Data Spaces (source online[21]).

The decentralized character is essential for the potential and distinctive character of Data Spaces. In data federation between a more limited and static group of companies or organizations, the relevant

[21] https://dataspaces.info/common-european-data-spaces

Funded by
the European Union

data is often duplicated and synchronized in a central location, for which the access rights of the parties involved are then regulated. This is neither feasible nor desirable from a scalability perspective. For example, an unnecessary amount of extra data storage would be needed, risk for data monopolies (as you already have on the internet) increases and these central sources could form a large-scale target for hackers. Data spaces avoid all these drawbacks. In essence, Data Spaces are an environment where suppliers and users of data can independently plug in and exchange data with each other. In addition, it is not necessary that they (always) interact directly with each other. Such flexibility and scalability will eventually make it possible to facilitate a range of new impactful applications based on data sources from various sectors. And for these reasons, a lot of technical, legal and economic efforts are currently being made to develop and roll out data spaces.



© International Data Spaces

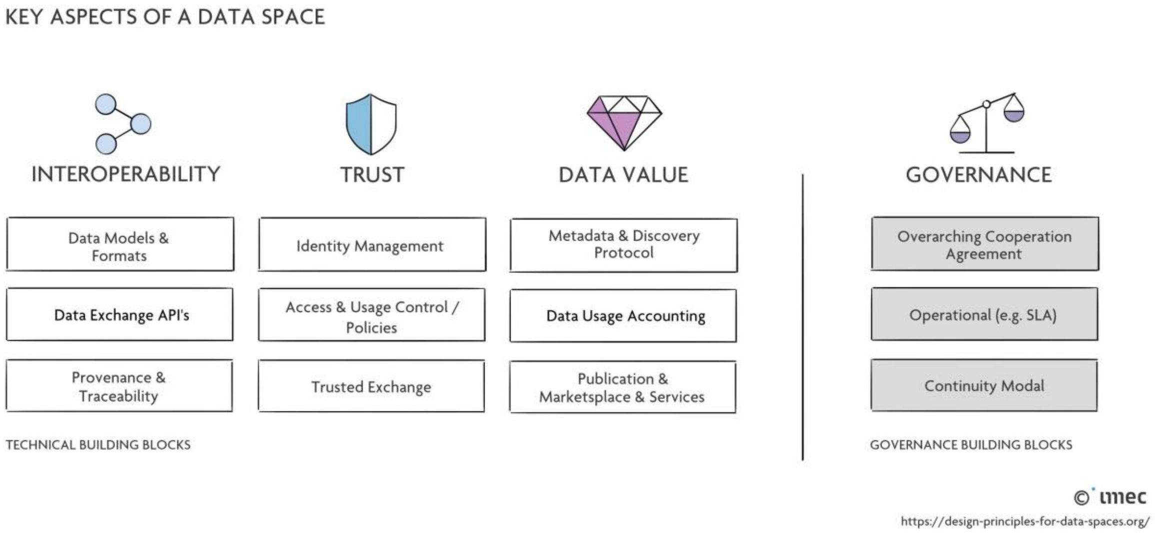
Fig 9. Schematic representation of the main actors, components and interactions in an IDS Data Space.

Among the existing European initiatives for Data Spaces is Gaia-X[22]. This consortium of industry, policy

[22] https://docs.gaia-x.eu/technical-committee/architecture-document/22.10/

Funded by the European Union

and research aims to have Data Spaces operational on a substantial scale in Europe by 2025. It maintains close ties with the International Data Spaces Association (IDSA), a global organization of more than 100 companies whose responsibilities include creating and deploying the International Data Spaces (IDS) reference architecture[23] (Figure 9). Since the end of 2022 - with financial support from the European Commission - there is also a Data Spaces Support Center (DSSC)[24]. This takes on a coordinating role in identifying the needs of the various sectors and coordinating existing initiatives. The organization also contributes to knowledge sharing about data spaces and also plays a key role in the development of European policy and regulations. Lastly, in September 2021, the Big Data Value Association, the FIWARE Foundation, Gaia-X and the IDS Association formed the Data Spaces Business Alliance [11]. The idea is to define a common reference technology framework, based on the technical convergence of existing architectures and models, leveraging each other's efforts on infrastructure and implementation.



KEY ASPECTS OF A DATA SPACE

INTEROPERABILITY	TRUST	DATA VALUE	GOVERNANCE
Data Models & Formats	Identity Management	Metadata & Discovery Protocol	Overarching Cooperation Agreement
Data Exchange API's	Access & Usage Control / Policies	Data Usage Accounting	Operational (e.g. SLA)
Provenance & Traceability	Trusted Exchange	Publication & Marketsplace & Services	Continuity Modal

TECHNICAL BUILDING BLOCKS

GOVERNANCE BUILDING BLOCKS

© imec
https://design-principles-for-data-spaces.org/

Fig 9. Key aspects of Data Spaces and how they are translated into (technical) solutions.

As shown in Figure 9, Trust is an essential concept for Data Spaces. Just as not just any app can access personal bank details, there must be a universal system for issuing licenses to reliable partners in the context of Data Spaces. Revoke of access must also be possible in the event of any abuse. Just as information on the internet can only be found because there are standards such as HTML and agreements such as domain names, Data Spaces also need widely supported standards and agreements. These start with the qualitative organization, filtering and semantic annotation of data, so that

[23] https://internationaldataspaces.org/publications/ids-ram/
[24] https://dssc.eu/

interested parties can find their way in it. For example, as proposed by the IDS Information Model vocabulary[25]. Subsequently, agreements are necessary that guarantee the findability of data and regulate their accessibility in function of the end user and the purpose they have with it. All this must be accompanied by the necessary economic and legal agreements that determine who owns the data and how they can determine to whom they are made available free of charge or for a fee. As soon as those answers become clearer, the technical infrastructure will also have to be equipped for this.

In the context of MareGraph, most data are completely open in accordance with the open data charter definition. However, a subset of the data contained within the EuroBIS dataset is restricted to be used for commercial purposes. This constitutes a scenario where existing Data Space concepts such as usage control policies[26] may be leveraged to comply with the legal requirements.

5 Reference Architecture for MareGraph

Considering the previously discussed technical concepts, we present a reference architecture (Figure 10) that aims to fulfill the requirements of cost-efficiency, flexibility and sustainability drawn in Chapter 1. The architecture definition is inspired from the architectural design proposed in [12]. It provides a high-level layered view based on the different functional roles that take place at different steps of the data production, publication and consumption processes. It also highlights different organizational roles that are relevant in the context of MareGraph, however these may be generalized as shown previously in

[25] https://w3id.org/idsa/core

[26] https://international-data-spaces-association.github.io/DataspaceConnector/Documentation/v5/UsageControl

Funded by
the European Union

Figure 7. Next, we provide a description of each layer.

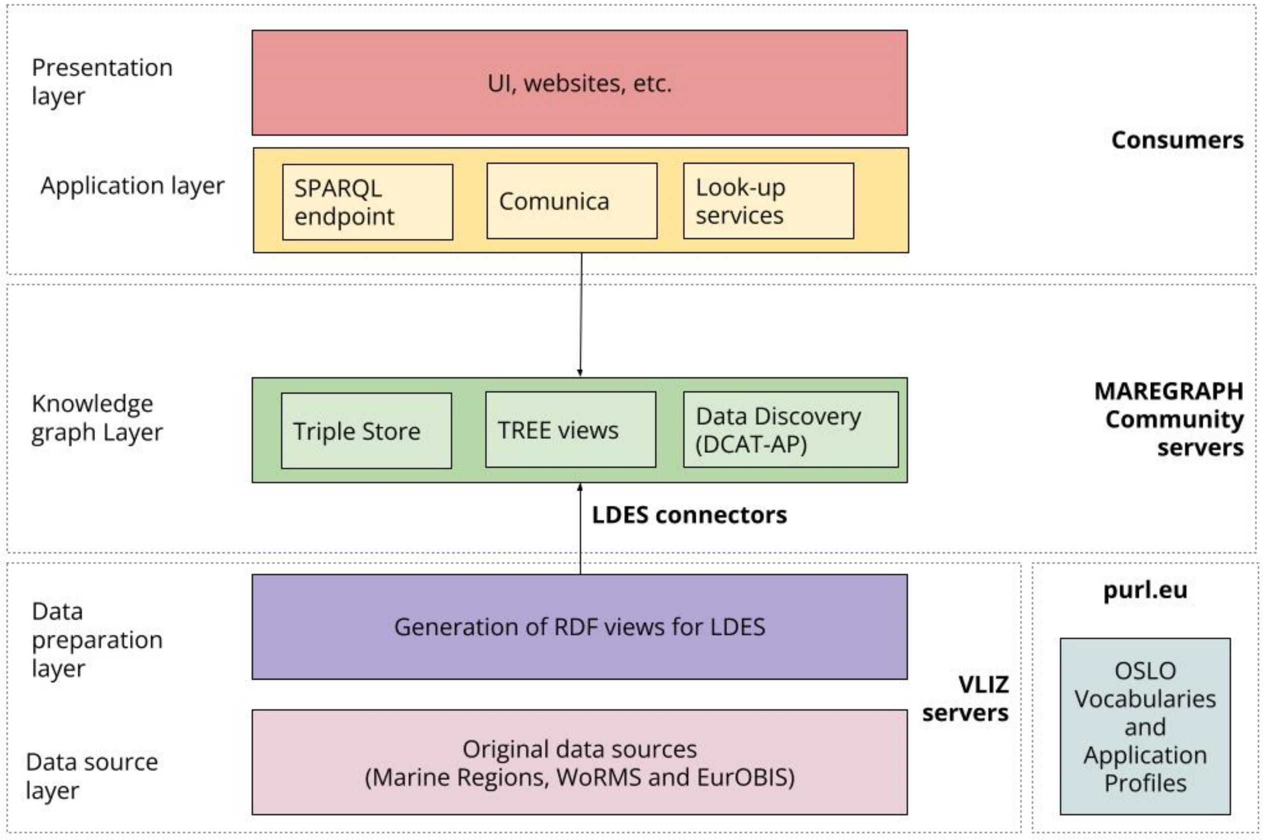


Fig 10. Reference architecture for the MareGraph project. The diagram shows both a layered view focused on functional roles and an organizational classification.

5.1 Data source layer

At the data source layer, we consider the original datasets considered for publication. The datasets may be already in the form of semantic data (RDF) or in other non-semantic forms, such as relational databases, among others. As previously mentioned, in the context of MareGraph we consider 3 main datasets:

1. **Marine Regions**: The Marine Regions system provides a digital gazetteer primarily focused on

Funded by the European Union

the marine environment. Originally targeted on the river Scheldt, and the Belgian part and southern bight of the North Sea, it now contains over 75k features worldwide. Entities in the Marine Regions gazetteer are characterized by their unique identifier, the Marine Regions Geographic Identifier (MRGID). This identifier allows users to unambiguously refer to a Marine Regions entity. Marine Regions entities can either have point, line or polygon geometries. This geometry is described by the derived centroid and (where available) bounding box coordinates. The Marine Regions dataset is already available as LOD and is even published by means of a LDES[27] [13].

2. **World Registry of Marine Species**: The aim of a World Register of Marine Species (WoRMS) is to provide an authoritative and comprehensive list of names of marine organisms, including information on synonymy. While the highest priority goes to valid names, other names in use are included so that this register can serve as a guide to interpret taxonomic literature. Within WoRMS, over 100 global, 12 regional and 4 thematic species databases are integrated with a common taxonomy. Currently WoRMS contains over 500,000 taxonomic names (from Kingdom to subspecies) [14]. WoRMS is supported on a relational database system called the Aphia platform.
3. **European Ocean Biodiversity Information System**: The European Ocean Biodiversity Information System – EurOBIS – is an online marine biogeographic database compiling data on all living marine creatures. The principle aims of EurOBIS are to centralize the largely scattered biogeographic data on marine species collected by European institutions and to make these data freely available and easily accessible. All data go through a number of quality control procedures before they are made available online, assuring a minimum level of quality necessary to put the data to good use. The available data are either collected within European marine waters or by European researchers and institutes outside Europe. The database focuses on taxonomy and occurrence records in space and time; All data are freely available online.

5.2 Data preparation layer

At the data preparation layer, the main objective is to ensure that the data is properly semantically annotated and structured according to the requirements of the LDES publishing approach.

Semantic annotation requires the specification of a data model in the form of a formal ontology, represented by the *OSLO Vocabularies and Application Profiles* component in Figure 10 that will be defined and published within the MareGraph project. However, specifying these resources is out of the scope of this paper. It is assumed that a semantic data model encompassing and conceptually integrating the 3 targeted datasets will be available when this architecture is implemented.

[27] http://marineregions.org/feed

Funded by the European Union

Publication in the form of a LDES, requires defining a versioning strategy and the associated SHACL shapes for data entities. The versioning strategy is related to the creation of unique URIs for every new version of an entity from each dataset, which will be published as an immutable object. An example from the already existing Marine Regions LDES can be seen in Figure 11. In this case, a URI strategy that appends a timestamp to the original entity (e.g., http://marineregions.org/mrgid/14**?t=1620659836**) was chosen. In this way, every new version of an entity is uniquely and persistently identified over time. The definition of SHACL shapes depends also on the availability of an ontology.

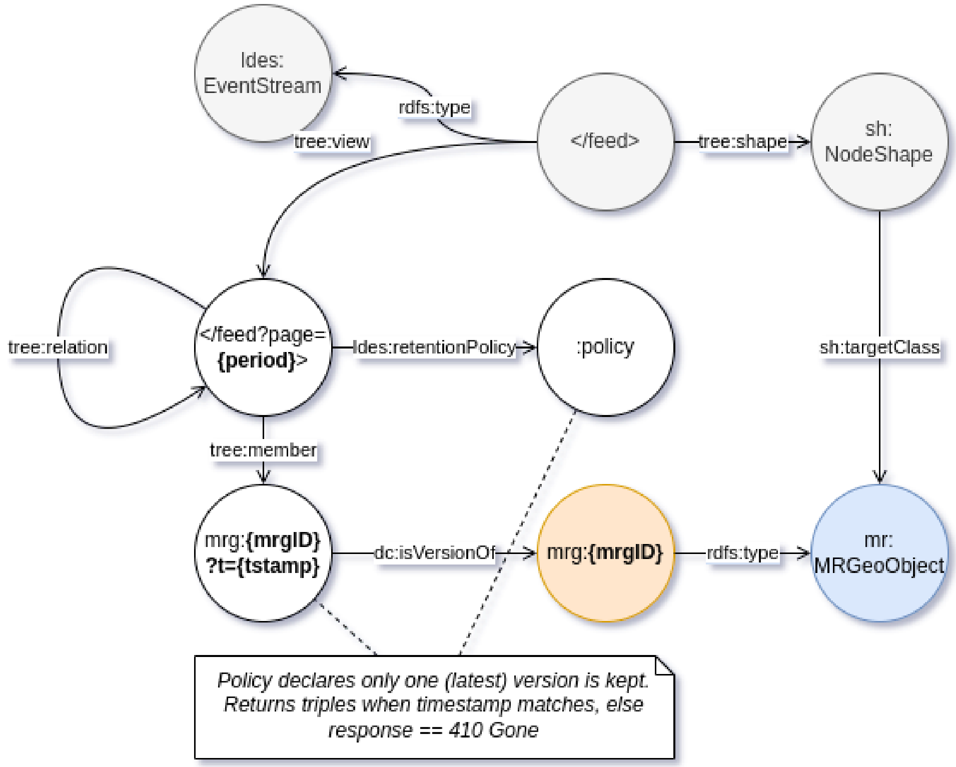


Fig 11. Marine Regions LDES data model.

For the publication of an LDES, a design decision can be made for including complete data entities within the LDES (i.e., data entities including all their properties and sub-properties), or to simply refer to the entity identifier. Depending on the design choice, a trade-off is established for data consumers. The former allows consumers to immediately access complete data entities with all their properties but may cause LDES generation and publishing performance issues for the data producer. Such is the case of the Marine Regions LDES, due to very large geometries associated with each marine region entity. The latter constitutes a lighter approach for the data publisher that may require less resources to produce and serve the LDES but at the same time, forces consumers to dereference each data entity when replicating and synchronizing with the LDES, which may cause performance issues. Compression techniques may be

Funded by
the European Union

explored at this layer, to enable more efficient data exchange between producer and consumers via LDES.

LDES generation approaches are no different from traditional RDF generation methods such as RML [15] and can be directly applied. Similar approaches could be followed to generate and publish corresponding LDES for WoRMS and EurOBIS datasets. It is envisioned that VLIZ, as data owner and producer, will host and maintain the different LDES.

5.3 Knowledge graph layer

The knowledge graph layer represents the materialization of an integrated knowledge graph. This layer relies on the existence of LDES feeds from the different data sources to materialize and keep in-sync not only an integrated knowledge graph hosted within triple store, but also additional TREE-based fragmentations targeted to support the commonly known use cases for these datasets (e.g., geospatial, and text search fragmentations). This entails the creation of a set of LDES connectors, able to read and synchronize from an LDES, and further process the data to materialize and fragment concrete versions of the different datasets.

Additionally, a metadata component based on the DCAT-AP specification is also envisioned. This component may be also derived from the different published LDES, by reading their included metadata which includes information regarding update times, type of data, among others.

The hosting of the triple store and the main TREE fragmentations are envisioned to be managed by the different partners of the MareGraph project. The LDES-based publishing approach facilitates this task since an LDES can be accessed over the Web and thus easily consumed.

5.4 Application layer

At the application layer we envision the creation and deployment of query agents that can support the creation of user-oriented applications. The meta-query engine Comunica [16] is mentioned given its adaptability to query heterogeneous interfaces such as SPARQL endpoints and various forms of LDFs, including TREE fragmentations. Further research is required to guarantee query performance from non-traditional interfaces, which will be carried out throughout the MareGraph project.

The hosting of these query agents is envisioned to be held by data consumers, given that they will be integrated as part of user-oriented applications *(presentation layer*) which will run, for example, in the browser of the users.

5.5 Presentation layer

The presentation layer includes the development of user-oriented Web applications that can provide the means to interact with the data, either by directly defining queries (e.g.,

Funded by the European Union

https://query.linkeddatafragments.org/) or through graphical interfaces that allow visualizing and interacting with the data.

6 Conclusion

In this paper we presented the conceptual background that motivated the MareGraph project. Starting from the legal framework defined by the EU Commission and the call for creating a better and more interoperable data ecosystem, of particularly targeted types of datasets. Despite dealing almost exclusively with fully open data, we also described the recent Data Spaces initiatives, as they constitute a more generalized framework for enabling interoperable and trustable data exchanges, that can guide the definition and implementation of a European Marine knowledge graph.

We also presented the technical background around Linked Open Data and more recent data publishing approaches that aim to overcome the data availability issues while guaranteeing maximum flexibility and cost-efficiency for data exchange processes.

We proposed a reference architecture that is centered around the concept of a Linked Data Event Stream, aiming to provide an enabling platform for data publishing and reuse. LDES, promises establishing dynamic data ecosystems where data providers are able to keep data publishing costs to a minimum and consumers are given full flexibility to reuse, transform and integrate data to support their particular needs.

Some technical challenges remain to be solved and will be investigated during the MareGraph project. Among these are, scalability of LDES generation and consumption and SPARQL querying performance of TREE fragmentations.

7 References

[1] Pronk, T.E., The Time Efficiency Gain in Sharing and Reuse of Research Data. Data Science Journal, 18(1), p.10. 2019.https://doi.org/10.5334/dsj-2019-010.

[2] Berners-Lee T. Linked data; 2006. Available from: http://www.w3.org/DesignIssues/LinkedData.html

[3] Bizer, C., Vidal, ME., Skaf-Molli, H. Linked Open Data. In: Liu, L., Özsu, M.T. (eds) Encyclopedia of Database Systems. Springer. 2018. https://doi.org/10.1007/978-1-4614-8265-9_80603

[4] Bizer, C., Heath, T., Berners-Lee, T., Linked Data – the story so far, International Journal on Semantic

Web and Information Systems 5 (3). 2009.

[5] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, Scientific American 284 (5), 34–43. 2001.

[6] I. Ermilov, M. Martin, J. Lehmann, S. Auer, Linked open data statistics: Collection and exploitation, in Knowledge Engineering and the Semantic Web, Vol. 394 of Communications in Computer and Information Science, Springer, pp. 242–249. 2013.

[7] C. Buil-Aranda, A. Hogan, J. Umbrich, P.-Y. Vandenbussche, SPARQL Web-querying infrastructure: Ready for action?, in Proceedings of the 12th International Semantic Web Conference, 2013.

[8] Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., Haesendonck, G., & Colpaert, P., Triple Pattern Fragments: a Low-cost Knowledge Graph Interface for the Web. Journal of Web Semantics, 37–38, 184–206. 2016.

[9] Colpaert, P., Building materializable querying interfaces with the TREE hypermedia specification, in Proceedings of the Managing the Evolution and Preservation of the Data Web (MEPDaW) workshop. 2022.

[10] Van Lancker, D., Colpaert, P., Delva, H., Van de Vyvere, B., Rojas, J. A., Dedecker, R., Michiels, P., Buyle, R., De Craene, A., & Verborgh, R. Proceedings of the 21th International Conference on Web Engineering. 2021.

[11] Data Spaces Business Alliance. Unleashing the Data Economy. Technical Convergence, 2022. Available from: https://data-spaces-business-alliance.eu/wp-content/uploads/dlm_uploads/Data-Spaces-Business-Alliance-Technical-Convergence-V2.pdf

[12] Anna Sofia Lippolis, Giorgia Lodi, & Andrea Giovanni Nuzzolese. Design of the technical services for knowledge graph management. Zenodo. 2021. https://doi.org/10.5281/zenodo.6685697

[13] Lonneville, B., Delva, H., Portier, M., Van Maldeghem, L., Schepers, L., Bakeev, D., Vanhoorne, B., Tyberghein, L., Colpaert, P. Publishing the Marine Regions Gazetteer as a Linked Data Event Stream. In 3rd International Workshop on Semantics for Biodiversity. 2021.

[14] Costello MJ, Bouchet P, Boxshall G, Fauchald K, Gordon D, Hoeksema BW, et al. (2013) Global Coordination and Standardisation in Marine Biodiversity through the World Register of Marine Species (WoRMS) and Related Databases. PLoS ONE 8(1): e51629. https://doi.org/10.137/journal.pone.0051629

[15] Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., & Van de Walle, R. RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In Proceedings of the 7th Workshop on Linked Data on the Web. 2014.

[16] Taelman, R., Van Herwegen, J., Vander Sande, M., & Verborgh, R. Comunica: a Modular SPARQL Query Engine for the Web. In Proceedings of the 17th International Semantic Web Conference. 2018.