Articles

MARISPACE-X: DIGITALIZING THE OCEAN

The Future of Maritime Data in the European Federated Data Infrastructure GAIA-X

By Daniel Wehner[1], Sergius Dell[2], Adrian J. Neumann[1] and Jann Wendt[1]

[1] north.io GmbH, 24118 Kiel, Germany
[2] TrueOcean GmbH, 24118 Kiel, Germany

Abstract

Within the Marispace-X project, a digital maritime data space based on data sovereignty, security, interoperability, and modularity according to the Gaia-X concept should be created. The article provides an overview of the envisioned Marispace-X digital infrastructure, a decentralized maritime data ecosystem, followed by exemplary applications for hydrographic surveys and some of the use cases defined in the Marispace-X project. These exemplary applications are related to data management, sharing and processing services that can be performed on a cloud infrastructure. This digital infrastructure should simplify data handling of all the different data types and formats that exist in the maritime domain.

Keywords: Data Management, Cloud Infrastructure, Maritime Data, Digital Ecosystem, Gaia-X, Decentralization, Open-source, Geoparquet

Résumé

Dans le cadre du projet Marispace-X, un espace de données maritimes numériques basé sur la souveraineté, la sécurité, l'interopérabilité et la modularité des données selon le concept Gaia-X devrait être créé. L'article donne un aperçu de l'infrastructure numérique Marispace-X envisagée, un écosystème de données maritimes décentralisé, suivi d'exemples d'applications pour les levés hydrographiques et de certains des cas d'usage définis dans le projet Marispace-X. Ces exemples d'applications sont liés à des services de gestion, de partage et de traitement de données qui peuvent être effectués sur une infrastructure cloud. Cette infrastructure numérique devrait simplifier la gestion de l'ensemble des différents types et formats de données existant dans le domaine maritime.

Mots-clés: Gestion des données, infrastructure cloud, données maritimes, écosystème numérique, Gaia-X, décentralisation, open-source, geoparquet

Resumen

Dentro del proyecto Marispace-X, se debería crear un espacio de datos digitales marítimos basado en soberanía, seguridad, interoperabilidad, y modularidad de los datos, según el concepto Gaia-X. Este artículo proporciona una descripción de los planes para la infraestructura digital Marispace-X, un ecosistema de datos marítimos descentralizados, seguido de ejemplos de aplicaciones para

levantamientos hidrográficos y algunos ejemplos de uso definidos en el proyecto Marispace-X. Estos ejemplos de aplicaciones están relacionados con los servicios de gestión, distribución y procesado de datos, que se pueden llevar a cabo en una infraestructura en la nube. Esta infraestructura digital debería simplificar la gestión de datos de todos los diferentes tipos y formatos que existen en el dominio marítimo.

Palabras clave: Gestión de Datos, Infraestructura en la Nube, Datos Marítimos, Ecosistema Digital, Gaia-X, Descentralización, Código Abierto, Geoparquet

https://doi.org/10.58440/ihr-28-a10

1. INTRODUCTION

The project Marispace-X is part of the European initiative Gaia-X. The idea of Gaia-X is to set standards for a digital ecosystem based on the values of openness, transparency, sovereignty, and interoperability as defined by the Gaia-X European Association for Data and Cloud AISBL (AISBL, 2022). Based on these standards federated services and data spaces are developed that can be used independently on various computing infrastructure, e.g., different cloud systems and vendors. These open and transparent standards will create an efficient system to share and collaborate on data while taking care of major topics like data sovereignty, security, and interoperability. While the concept of Gaia-X could be applied to any field, Marispace-X deals explicitly with data in the maritime domain. As the ocean is a challenging environment considering costly data acquisition and complex data transfers, a standardized digital ecosystem can increase the efficiency of the whole data value chain from acquisition to complex analytical data products. The data acquired in the marine area may include hydroacoustic, satellite, optical and magnetic surveys, or chemical, biological, and physical measurements of the environment. Therefore, several data types and formats exist that are used within the maritime domain. However, for the interpretation of the data, different data types and formats often need to be evaluated together and not separately. The digital cloud infrastructure developed in Marispace-X should simplify the handling of these different data formats and types.

The Marispace-X project started in January 2022 and is funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) for three years, until the end of 2024 with a total budget of 15 million Euros. The goal is to establish a legal entity around Marispace-X and a solid funding mechanism to ensure the long-term operation of the data space, its governance, and a worldwide adoption. The first part of the project deals with the standards and setup for the underlying technological architecture, federated services, and data spaces. The second part considers the specific use cases Internet of Underwater Things, Offshore Wind, Munitions in the Sea and Biological Climate Protection. For the use cases numerous stakeholders from industry, academia and governmental institutions are involved to find solutions for their specific problems with marine datasets. Here, we give an overview of Gaia-X (Section 2) and describe in more detail the Marispace-X project (Sections 3 and 4) and potential applications to hydrographic data (Sections 5 and 6). Finally, an exemplary outlook for some of the planned use cases in Marispace-X is given (Section 7). The intent for the article is to start from the very general overview of the Gaia-X data infrastructure and end with some specific implications for the marine environment.

2. GAIA-X: OVERVIEW

To set the standards for the Gaia-X digital ecosystem based on the desired European values, as many stakeholders as possible from different fields should be involved. Therefore, the standards are developed by the Gaia-X European Association for Data and Cloud AISBL which was established in September 2020 and is composed of companies, research institutions, associations, governmental and political institutions from all over the world. The association is open for more institutions that like to join the initiative. As subsets of the European Association for Data and Cloud AISBL there are national Gaia-X Hubs that represent the partners and projects in the respective country. For instance, in the Gaia-X Hub Germany the projects are separated into sectors within the working groups: Agriculture, Energy, Finance, Geoinformation, Health, Industry 4.0/SME, Mobility, Public Sector, Smart City/Smart Region and Smart Living. The Marispace-X project is part of the Geoinformation domain while it exclusively deals with data in the marine environment. Besides the European association and the national hubs, a Gaia-X community is established where everyone can participate and exchange ideas.

From a technical perspective Gaia-X consists of four core elements: standards, data spaces, federated services, and business services:

- Standards: The Gaia-X standards are based solely on existing data and sovereignty standards, as well as related infrastructure components, e.g., based on the International Data

Spaces (IDS) Reference Architecture Model (RAM) (Gaia-X, 2022; IDSA, 2019; IDSA, 2021).

- Data spaces: The data space concept is that data is stored locally rather than centrally and is only exchanged when needed. The data spaces should all follow the same standards for storage and sharing which allows data interaction between trusted partners.
- Federated services: Federated services are responsible for the interchange and processing of cloud data and all relevant implications. For instance, data owners can share information (metadata) about existing data between them using federated services, reducing unnecessary data transfers, and utilizing the federated services only for the desired data.
- Business services: Business services are standard procedures and policy rules defined on the business level. By establishing uniform principles for collaboration inside and across industries, it should allow to generate new prospects for value creation and business innovation.

3. MARISPACE-X: OVERVIEW

The Marispace-X project is funded by the German Federal Ministry of Economics and Climate Protection (BMWK) for three years, starting in January 2022. Its objective is to create a digital maritime data space based on data sovereignty, security, interoperability, and modularity according to the concept of Gaia-X. It will provide new ways in maritime big data processing and analysis of sensor data across edge, fog, and cloud computing. A concrete example of these three computing components for a ship-based multibeam survey could be as follows. First processing steps are directly performed on the multibeam sensor (edge). The acquired data is than preprocessed on local computers on the vessel (fog), while the algorithms for the computation could be retrieved from the cloud. Then the preprocessed data is further analyzed and shared in the cloud.

The newly build maritime data ecosystem will allow stakeholders from business, science, public authorities, and non-governmental organizations (NGOs) to securely manage, share, and analyze data acquired about and from the ocean. This efficient system to share and analyze all available data will create new insights about the ocean, will allow information-based decisions to be made, and will lead to new developments of federated services in the future.

The project consortium is led by cloud provider IONOS SE and software developer north.io GmbH. Consortium partners include the maritime big data specialist TrueOcean GmbH, the Fraunhofer Institute for Computer Graphics Research (IGD), GEOMAR – Helmholtz Centre for Ocean Research, Stackable GmbH, MacArtney Germany GmbH, Kiel University, and the University of Rostock. The involvement of additional national and international associated industry partners and maritime stakeholders ensures industry and application-oriented development in all areas.

Marispace-X includes four concrete use cases that leverage upscaled storage, processing, and analysis of maritime data and addresses several key challenges of the next decades, such as climate change, marine conservation, and digital transformation. The four use cases are briefly described below.

3.1 Internet of Underwater Things

Data-driven underwater technologies and sensor networks are an important tool of the digitalized Blue Economy of the future. This use case aims to enable sensor technology to capture a digital twin of the underwater world in real time. At the same time, Gaia-X compatible infrastructures are developed for data acquisition and processing from edge (the sensor) to the cloud. This is done through a feasibility study based on the Ocean Technology Campus of Fraunhofer IGD and through experimentally advancing the Internet of Underwater Things by creating a Gaia-X compli-

ant data environment. Processing and analysis tools should be developed for the edge components (sensors) to reduce the amount of required data that needs to be transferred to the cloud.

3.2 Offshore Wind

Offshore wind farms are a key to lower CO_2 emissions in the energy sector and Marispace-X helps to develop solutions for collaborative data collection, management, and smart asset handling. The use case's goal is to develop data-driven applications and generate trustworthy data spaces for collaboration and efficiency gains of offshore wind farms. This should be achieved through a collaborative development of tools to map the entire process chain in the collection and analysis of data and by enabling end-users to store, manage and analyze data over the full lifetime of projects (25 years +). A cloud tool to simplify the handling of large amounts of heterogeneous data sources would benefit all involved stakeholders.

3.3 Munitions in the Sea

It is estimated that German waters alone contain about 1.6 million tons of old munitions (Böttcher et al., 2011). Therefore, efficient ways of data management and analysis tools to tackle the problem are needed. The use case aims for improvements in the management and analysis of data for munitions disposal through new developments in cloud, fog, and edge computing. This includes data management and visualization of heterogeneous data sources (e.g., historic documents, hydrographic survey data, underwater photos) in the cloud to support decision making on prioritized areas for munitions clearance. In addition, data quality measures should be implemented to analyze the data and conclude whether it is feasible to detect certain munition objects in the available data. These analysis tools may be implemented on edge, fog, or cloud components. Another important issue is the data sharing in the security-critical ordnance domain that should be solved by the Gaia-X compliant implementation of federated identity and trust services.

3.4 Biological Climate Protection

Ocean plants have high CO_2 storage capacities. The use case is intended to provide a basis for determining and optimizing the CO_2 savings potential of macrophytes. The goals of the use case are to identify, quantify and optimize the potentials of using CO_2 storage capacities of the oceans in the fight against climate change. This is done by establishing novel analysis methods like sensor fusion of satellite remote sensing data and underwater acoustic data. Further potential analysis tools could be related to the development of AI-based predictors for optimal macrophyte settlement areas (Held and Schneider von Deimling, 2019). The use case could provide efficient methodologies to map coastal areas and indicate areas with existing macrophytes, potential settlement areas, and unfeasible areas for macrophyte settlements.

4. MARISPACE-X: INFRASTRUCTURE, FEDERATED SERVICES AND DATA SPACES

The Marispace-X project is based on the guiding principles of Gaia-X and addresses both the focus on data spaces and federated services. The focus is on interoperability and portability of data and services under the condition of transparency and sovereignty. Marispace-X is developing a digital collaborative data space including various federated services that form the basis for future Gaia-X-compliant developments. Various federated services will be implemented in the areas of sensor data acquisition (edge), the various pre-processing computing steps on local machines (fog), and the final analysis of the data in the cloud-based Software-as-a-Service (SaaS) platform. Federated services enable the connection between different components (e.g., edge/fog/cloud, different cloud storages), allowing data from different sources to be managed, shared, and pro-

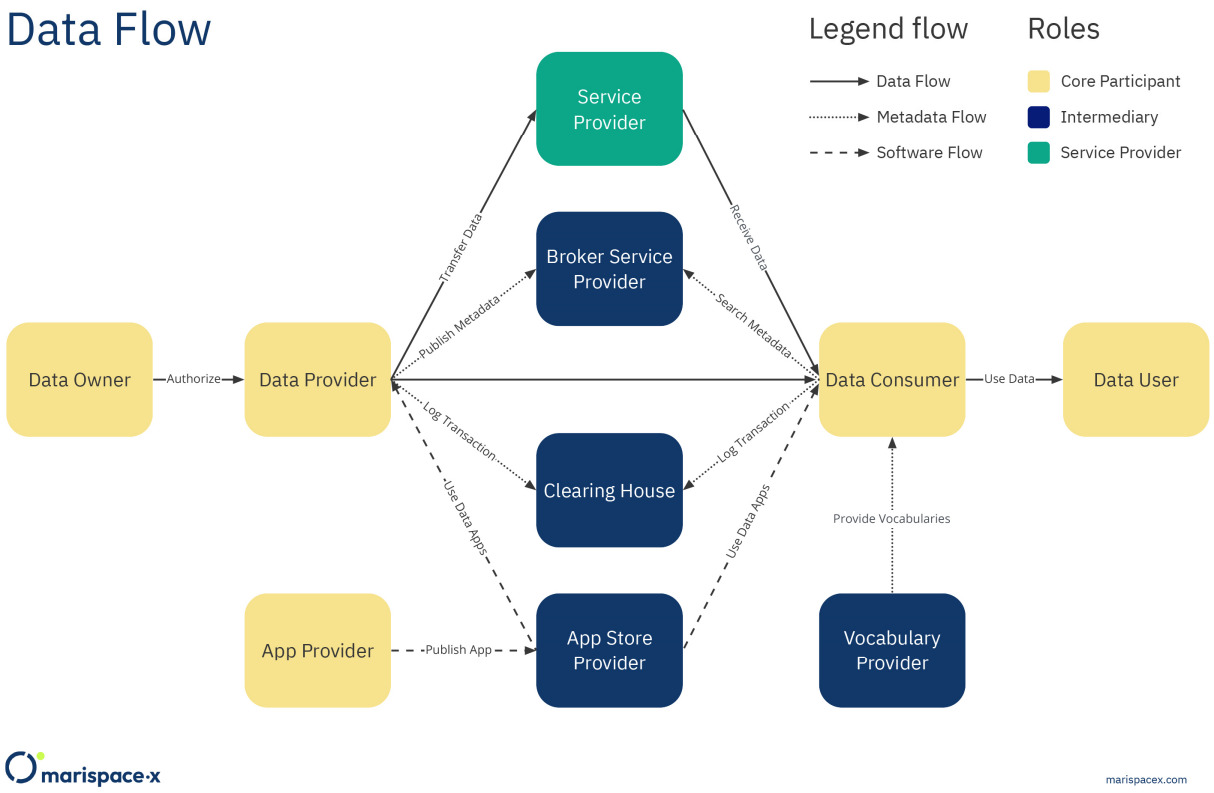

Figure 1. *Sketch of involved roles Core Participant, Intermediary, and Service Provider for the data flow in the IDS reference architecture model, version RAM 3.0. In addition, the flow of data, metadata, and used software components is illustrated (adopted from IDSA (2019, 2021)).*

cessed using the same services. This is possible if the federated services follow the defined specifications (Gaia-X, 2022). A general description of federated architectures for data management can be found in Heimbigner and McLeod (1985). The overarching goals of this infrastructure are advancing the digitalization of the maritime domain, realizing maximum efficiency between the components (edge/fog/cloud), and using the possibilities of interoperability, portability, and identity & trust in the context of Gaia-X in an application-oriented manner. The technical foundation of Gaia-X is comprised of the IDS reference architecture model (IDSA, 2021) and the federated services developed by the Gaia-X AISBL. Here, we explain the infrastructure, federated services, and data spaces based on the IDS reference architecture model version RAM 3.0 (IDSA, 2021), as this is the latest officially approved version while writing this article. The roles for this infrastructure as the Core Participant, Intermediary and Service Provider are described below (***Figure 1***). However, as these standards are still evolving and a new version, version RAM 4.0, is planned to be published in 2022 (IDSA, 2022), changes to the infrastructure components described below are still possible. The newer architecture is planned to be more decentralized, where the connector (***Figure 2***) is the main component, compared to version RAM 3.0. For another example, about the role of the connector and data spaces, interested readers are referred to Totzler and Tschabuschnig (2022).

Every time data is shared, Core Participants are involved and necessary in the provider/consumer model. These Core Participants are Data Owner, Data Provider, Data Consumer, Data User, and App Provider (***Figure 1***). Any organization that owns, wants to provide, and/or wants to consume or use data can play the role of a Core Participant.

A layer of several Intermediaries operates as trusted entities in the provider/consumer model. These Intermediaries are the Broker Service Provider, Clearing House, Identity Provider, App Store Provider, and Vocabulary Provider (***Figures 1*** and ***2***). All these Intermediaries provide federated services. The interaction and operation of the Intermediaries are facilitated by the connect-

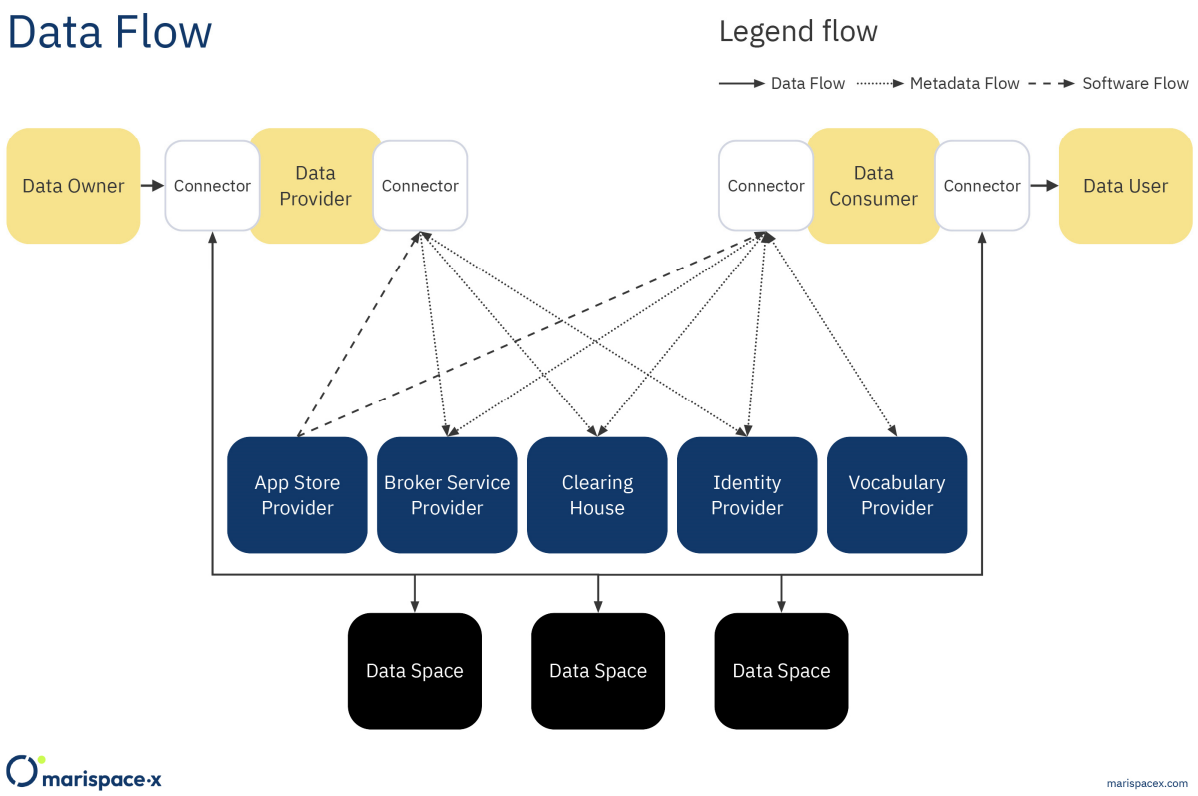

Figure 2: *Sketch of Core Participants and Intermediaries and their interaction with the connector and data space. The connector links the data spaces between the Data Owner and Data User, where the metadata and access rights are checked through the Data Provider/Data Consumer.*

or, the central technical component of the data space architecture (***Figure 2***). Only trusted organizations are allowed to take on these responsibilities. These regulations benefit members by establishing trust, providing metadata, and developing a business model around their services. The data space is linked via the connector between the Data Owner and Data User, where the metadata and access rights are checked through the Data Provider/Data Consumer (***Figure 2***).

The Service Provider is considered an organization that can provide the technical infrastructure for Core Participants that do not have the IT infrastructure themselves. The Service Provider could also only provide additional services to process or analyze the data if the Core Participant has the infrastructure available internally.

The different Core Participants and Intermediaries (***Figures 1*** and ***2***) are briefly explained in the following.

A Data Owner is defined as a legal entity or natural person who creates or owns the data and can set the data usage policies, like access rights for other entities (IDSA, 2019). The Data Provider supplies the technical components to exchange data between Data Owners and Data Consumers. For instance, the Data Provider can submit metadata to the Broker Service Provider and store logging information about the data transfer at a Clearing House. The Data Provider could often be the same participant as the Data Owner. The Data Consumer receives the data and hence is the mirror entity of the Data Provider. The Data Consumer could receive the metadata from the Broker Service Provider. The Data User is the legal entity that has the right to use the data of a Data Owner as stipulated in the usage policy. Also, the Data User and Data Consumer could be the same participant. In addition, an App Provider creates apps as federated services that can be used from all Core Participants for data handling, management, sharing, and processing.

The Broker Service Provider is an Intermediary that maintains and manages data sources.

Hence, the major task is to receive metadata from the Data Provider and provide metadata to the Data Consumer. The Clearing House keeps track of all activities that occur during the data exchange by logging the information. This information might include error logging or billing information. The Identity Provider offers a service that allows the Core Participants to create, maintain, administer, monitor, and authenticate their credentials. This is critical for the secure data handling and the prevention of unwanted data access. The App Store Provider makes the apps from the App Provider available and provides application programming interfaces (APIs) for publishing and retrieving the apps, as well as for corresponding metadata. The Vocabulary Provider maintains and stores vocabularies that are required to annotate and describe the data. This is provided through an information model that serves as the foundation for data source descriptions.

5. MARISPACE-X: CLOUD-BASED OPEN-FILE FORMAT

In the digital ecosystem of Marispace-X a unique, cloud-based and open-file format should be used. Data acquired during hydrographic surveys is written to digital files for further processing, sharing, analysis and long-term storage. This is usually achieved by using specific file formats, which might be proprietary or open-source and depend on the recording system, acquisition type, and sensor model. Data acquired by side-scan sonars, for example, is conventionally stored using the *XTF* binary format introduced by Triton in 1988 (Triton Imaging, 2016). Data recorded by sub-bottom profilers is conventionally stored using the *SEG-Y* binary format, originally developed in 1975 to save single-line digital data on magnetic tapes (SEG, 2017). Data acquired by single or multibeam echosounders is usually stored using proprietary binary formats, e.g., *ALL f*ormat by Kongsberg or *S7K* format by Teledyne Reson, which were also developed in the past century (Kongsberg, 2018; Teledyne Reson, 2019). Further open-file formats commonly used to store multibeam data are the binary format *GSF* (Ferguson and Chayes, 2009; Leidos, 2019) or row-based ASCII formats, e.g., *CSV*, *XYZ* (RFC, 2005*)*. While ASCII formats can be easily accessed by almost every software, the format is not efficient for storing, accessing, and sharing large amounts of data. In contrast, every binary format is designed to best store the specific sensor measurement. However, to access and merge data of different binary formats is more challenging, although all of them have a similar file structure and byte stream organization. The binary formats start with a file-header block. It provides general information applicable to all data streams in the file (data streams are, e.g., water depth, x-coordinate, y-coordinate, etc.). Then data packets, also called records, datagrams, or messages, follow the header block. These packets contain data related to a single measurement. It can include amplitudes of the backscattered signal as well as detailed information about the geographic location, and physical characteristics of the recorded signal. Both data packets and file-header blocks are normalized and described in corresponding format specifications. After data packets, additional file data trailers might follow. They usually contain some additional information about the survey or meta information about the file itself. Since the initial format releases, however, there have been significant advancements in data acquisition, such as three-dimensional techniques and high-speed, high-capacity recording. Hence, the amount of data is steadily increasing and many of the conventionally used data formats (binary or row-based ASCII) are less suitable for processing, sharing, and storing in the modern cloud environment.

We propose to store the different sensor measurements in datasets, which fulfill the following criteria:

- a dataset contains self-describing syntax, so-called data schema.
- its data blocks can efficiently be extracted so that data are retrieved on a low-latency basis; this includes available application programming interfaces (APIs) to most programming languages.
- it is cloud-native, i.e., data is accessible across different platforms, inter-operable and governed by global standards.
- it is an open-file format, so that it can be implement into internal and external workflows.

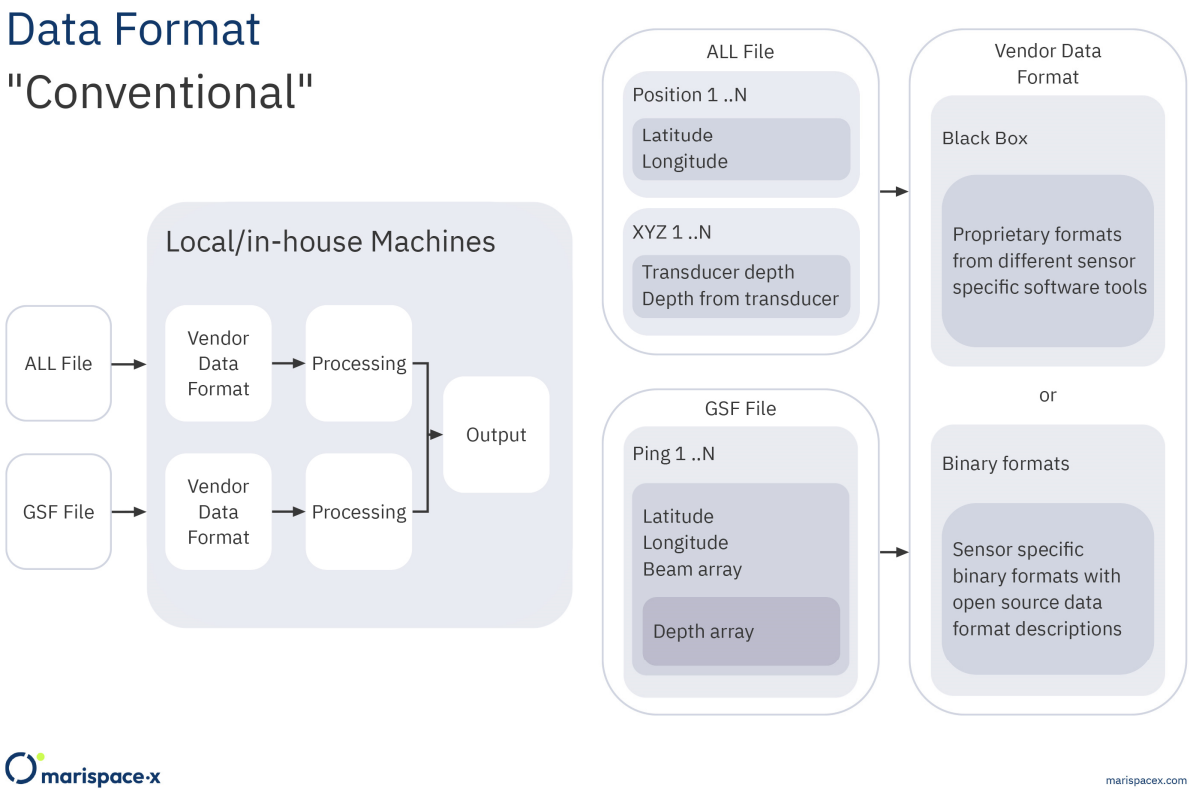

Figure 3. *Sketch that illustrates the conventional data processing workflow on local machines, often using proprietary data formats as the vendor data format. The output could be, for instance, a bathymetric grid of the input data files (ALL and GSF).*

We suggest using the Apache Parquet (*Parquet*) format for storing the sensor measurements. *Parquet* represents a column-oriented data file format maintained by the Apache Foundation. It is an open-file format and is actively being updated and maintained. *Parquet* is built from the ground up with complex nested data structures in mind and therefore it is superior to simple flattening of nested name spaces (Apache Parquet, 2022). *Parquet* supports very efficient compression and encoding schemes. A *Parquet* dataset is an object, which implies it is suitable for storing data in any object storage, e.g., IONOS cloud, AWS Cloud, Microsoft Azure. Simply speaking, it combines the advantages of ASCII formats (easily accessible) and binary formats (efficiently compressed), while providing fast data access. In addition, a *GeoParquet* format is in development (de la Torre, 2022) that extends the *Parquet* format, e.g., through the integration of geometry data types (points, lines, polygons). If the *GeoParquet* format will be further developed it might be a suitable alternative for storing geospatial data. A slightly different approach to handle all different data formats would be to write a generalized metadata file for available formats as presented by Calder and Masetti (2015). This would simplify the access of the different file formats, while it would not allow a more efficient data processing in the cloud as enabled by an optimized data format like the suggested *Parquet* file format.

A *Parquet* record is organized as a collection of row groups, where a row group consists of a data block for each consecutive column, and each data block consists of one or more pages of column data. That is, a single row group contains data for all columns for a given number of rows (***Figure 4***, Parquet Data Format). Each *Parquet* file has the footer, which contains metadata information among other column statistics. The column statistics are available for the row group and include minimum and maximum values. This allows the reading of the entire *Parquet* file to be skipped if querying a data block of interest. The *Parquet* format also allows for storing nested data structures, which implies it can be used to store point cloud data, vector data, raster data, search indices, etc.

The dataset concept allows the ability to directly work with the data streams. That is instead of

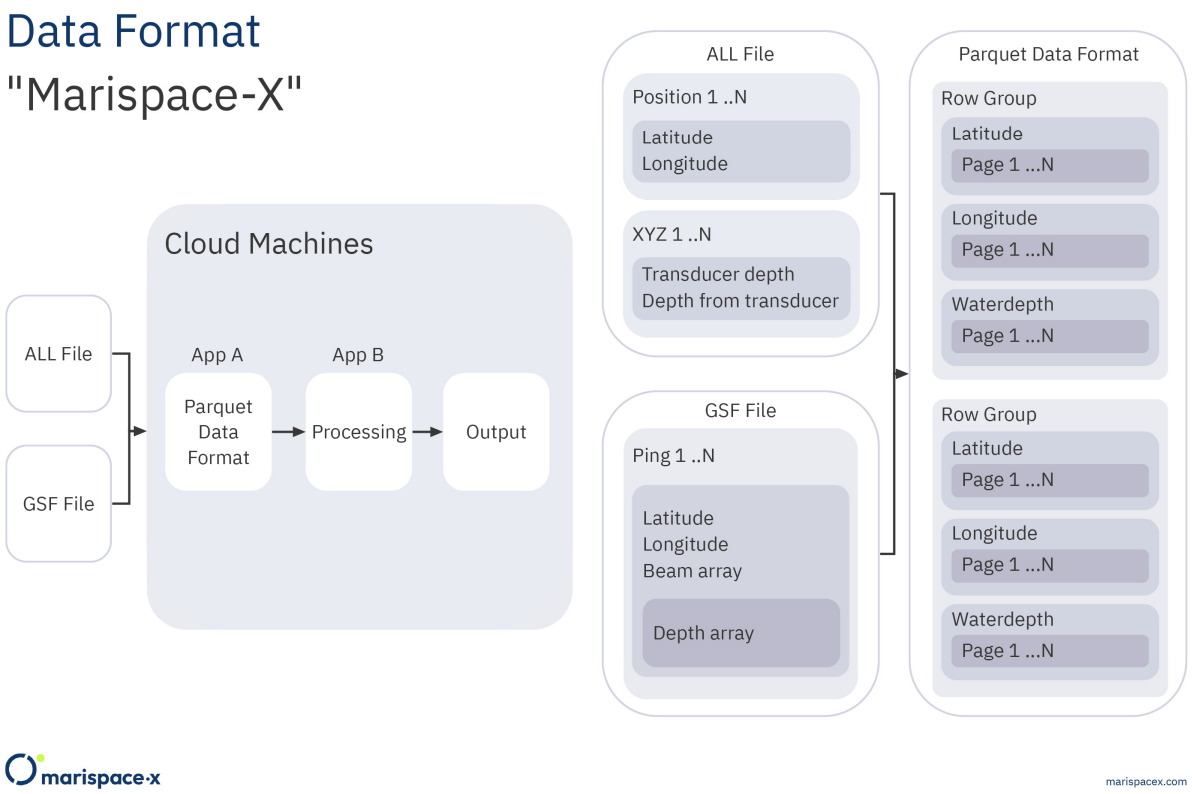

Figure 4. *Sketch that illustrates the Marispace-X data processing workflow on cloud machines using the open-file data format Parquet. The output could be, for instance, a bathymetric grid of the input data files (ALL and GSF).*

working with the files as static entities, we work dynamically by identifying and extracting data streams and combining them to a new dataset. Note, the data streams can arbitrarily be combined depending on the requested processing. They also can come from different input file formats and sensors. After the data processing is finished the datasets can be deleted to free resources or be stored in the cloud for further usage or reuse. ***Figures 3*** and ***4*** illustrate our approach for multibeam measurements. Here, a use case is considered where datasets from two different hydrographic surveys are combined to calculate an output result, which could be a bathymetric grid. The measurement files are stored in two file formats, e.g., *ALL* and *GSF,* and the files are saved on two different storages. Conventionally, the measurement files would be processed in a format-specific manner. First, the position records are evaluated for the *ALL* format and the coordinates (latitude, longitude) are processed and merged with the depth values from the XYZ package. Then, for the GSF format, the coordinates (latitude, longitude) from the ping records are processed and merged with the depth array values. Then, the intermediate results from both processing steps are merged and used to calculate the output, e.g., a bathymetric grid (***Figure 3***). Using the approach in Marispace-X, the measurement files from the different storages (data spaces) are uploaded via a dedicated app *A* (federated service) from the app store (***Figure 4***). This app *A* only extracts the required data streams (e.g., coordinates and water depth), converts the data to a single *Parquet* dataset and passes it to another dedicated app *B* for computing the output, e.g., the bathymetric grid (***Figure 4***). This also has the benefit that all processing steps can easily be performed and repeated on the single *Parquet* file instead of conducting these operations separately on the different binary formats.

6. MARISPACE-X: CLOUD COMPUTING IN PRACTICE

An exemplary application of federated services in a cloud infrastructure is discussed below. In general, cloud computing is defined by the National Institute of Standards and Technology (NIST)

as a model which enables ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources among other networks, servers, storage, applications, and services (Mell and Grance, 2011). The cloud infrastructure provides all the servers that client applications need to keep running at scale, while the servers are automatically updated accordingly. Computational resources are allocated dynamically, released with minimal management effort, and are instantly auto scaled to meet elastic demands for the required tasks and computations. The scaling can happen horizontally, referred to in/out scaling, and vertically, referred to up/down scaling. Horizontal scaling means that more physical servers are allocated to perform the required tasks and computations. Vertical scaling means that more capacities or more powerful single machines are allocated to perform the required tasks and computations. The tools that perform the tasks and computations on the data are defined as services or microservices. In the Gaia-X concept these specific tools are provided as federated services through the App Store Provider or Service Provider (***Figures 1*** and ***2***) and hence they are available for the client's specific needs on request. These federated services allow a client to use available resources, instead of building the infrastructure and every service on his own, e.g., low-level services for the management of large datasets or other data analysis services. Moreover, the established practice of micro-billing for services allows cost savings by charging only for the compute resources and duration the code uses to execute the workflows defined by the client. This information is tracked by the Clearing House (***Figure 1***). A similar approach is presented by Wright and Wright (2017) that explicitly deals with hydrographic data, while the Marispace-X project should also include other maritime data sets as magnetic, optic or satellite measurements.

As an example, we present an application called data service, which is suitable for importing and analyzing large volumes of high velocity data from various sensor sources stored as structured binary files or unstructured ASCII files like *CSV* files (Section 5). The data service architecture follows a microservice pattern. A microservice is a loosely coupled, independently deployable unit of code often packed into a container, for example a docker container. Microservices typically communicate with lightweight mechanisms, often a hypertext transfer protocol (HTTP) resource API and are built around business capabilities (Lewis and Fowler, 2014). The services packed into containers are independently deployable on different cloud infrastructures (clusters), which means the services are cloud agnostic. The deployment is fully automated by a deployment system, for example Kubernetes. Kubernetes is an open-source orchestration tool developed by Google and now run by the Cloud Native Computing Foundation. Kubernetes provides a highly resilient infrastructure including automatic deployment, rollback, self-healing, and scaling of containers based on the required CPU usage. It also provides dedicated endpoints to a microservice, which allows a load balancing of network traffic. ***Figure 5*** provides a simple sketch of the exemplary data service deployed on a cluster. From the client device, which could be a mobile phone, desktop tool, or through a web browser, several requests are sent to the data service, which runs in the cluster, generating network traffic. The cluster provides a controller and load balancer that distributes instances of the data service on different pods to efficiently scale the workload on the cluster. An exemplary request could be that files of different data formats (*GSF, ALL*) from the same sensor type, e.g., a multibeam echosounder, could be uploaded, statistics could be computed, and the data could be converted to the unified *Parquet* format to efficiently store and share the data. The requests are first processed by the controller on the cluster and passed to a load balancer, which distributes the traffic across the pods, the elementary unit of the cluster. Every pod contains an instance of the data service, which means it operates independently from the other pods. The data service performs tasks like the validation of the client's request, gets the files from the external storage (data spaces), computes basic statistics on the data, converts the files to a unified dataset (*Parquet*, Section 5), and writes datasets to external storage (data spaces). The external storage could be one or several different instances (***Figure 5***). The number of pods can automatically be increased or decreased by horizontal and vertical scaling. This scaling is done automatically through the cluster orchestration system, e.g., Kubernetes. This allows to load and convert datasets from large surveys in a relatively short time and prevent cluster instances sitting idle after the survey load is finished, which would only waste computing resources. From the business level the cluster is an Infrastructure as a Service (IaaS) and the

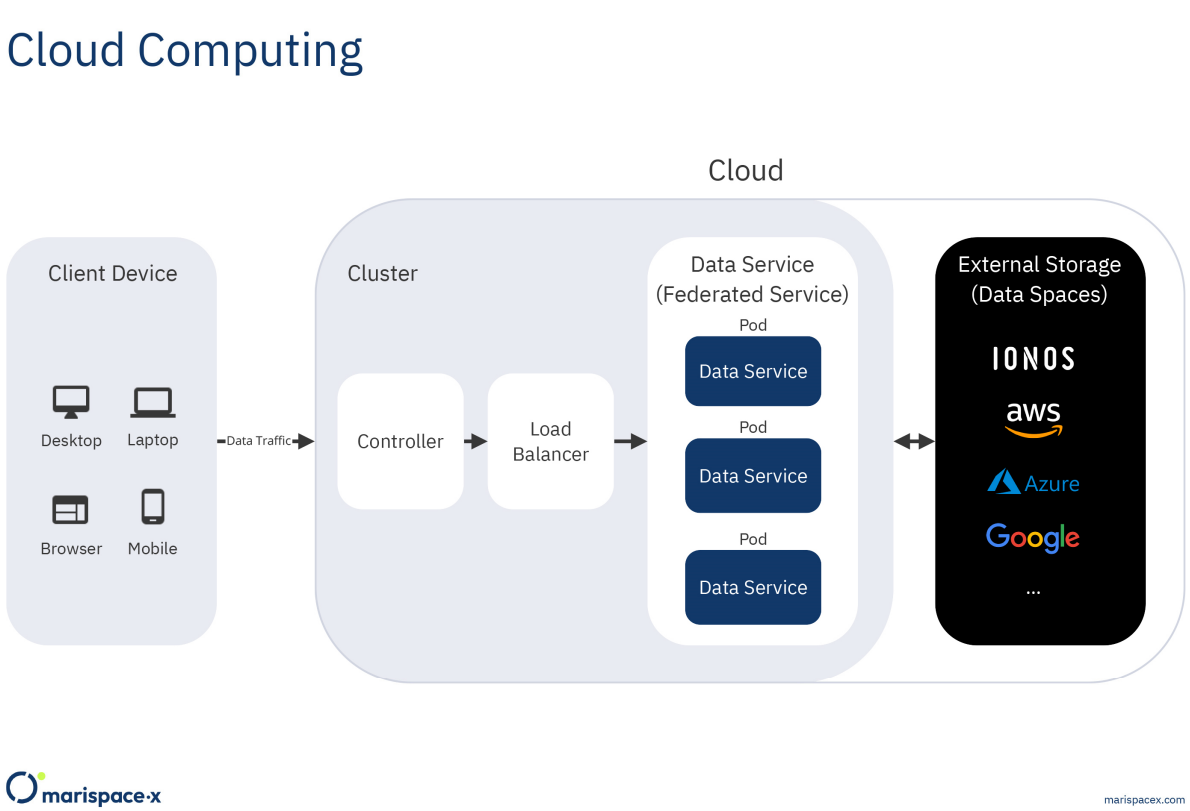

Figure 5. *Sketch of cloud infrastructure and deployed exemplary federated service (data service), which is part of the Intermediaries (**Figures 1** and **2**). The data service is started by the user via a client device and depending on the required workload, several instances of the data service run on the cluster as assigned by the load balancer. The exemplary data service performs tasks like importing, processing, and converting different sensor data formats into the unified, cloud native, open-file data format Parquet. The data is retrieved from and stored on the external storage, where a few examples of cloud storages are given. However, within the project the external storages need to comply with the Gaia-X standards.*

federated service is a Software as a Service (SaaS). Hence, the client (1) could have his own cluster and only needs the federated service from a provider, (2) could request the cluster and service from one provider or (3) from two different providers.

In Marispace-X the underlying infrastructure and services are developed and in a second step different services should be tested for the four use cases. To get a first impression of the scaling to efficiently import and convert binary sensor data into the *Parquet* format, a computation example is given here. We conduct our tests, to convert *GSF* files to *Parquet* files, on three file sets as described in ***Table 1***. All test sets contain roughly the same number of sounding points (7.000.000.000), but they differ in the individual file size and number of files. We also modify the number of extracted data streams for a file set that are converted to *Parquet*: four (x, y, water depth, intensity) and eight (x, y, water depth, intensity, across track (y), along track (x), heave, heading). The different file sets allow us to compare the efficiency using varying amounts of available pods, and the efficiency to load many small files or fewer big files with the same amount of data (same number of soundings). The hardware system used for the test has 32 GB of random-access memory (RAM), 8 central processing units (CPUs) where each processor is an Intel(R) Xeon(R) Gold 6348 @ 2.60 GHz. The maximum bandwidth during the test was about 1 Gbit/s.

The resources and time required for the upload and data service (data format conversion) are displayed in ***Table 2***. The upload time is always the same for each file set as the input data does not change. The conversion time reduces with the number of pods. As we directly send the payload to the dedicated endpoint in the cluster, which solely distributes the whole load, the conversion time is not reduced by the number of nodes exactly. That also explains a moderate decrease in the conversion time for many small files (file set A). In addition, we see a slight increase in the

conversion time for a larger number of data streams. This is expected as every pod was limited by single CPU, i.e., we did not perform any parallel processing during the scan, data stream retrieval and storage of the *Parquet* output.

Table 1. *Three different data sets (A, B, C), used for the computation example, with different amounts of files and file sizes, but each data set contains roughly the same number of sounding points (7.000.000.000).*

File set	Number of files	Size of single file [MB]	Number of pings in single file	Number of beams per ping	Number of soundings in single file
A	8000	14,042	1701	512	870.912
B	2055	54,507	6622	512	3.390.464
C	584	194,781	23307	512	11.933.184

Table 2. *Computation example to illustrate the time required to upload and convert different GSF files containing multibeam data into the open-file data format Parquet. Number of data streams refers to 4 (x, y, water depth, intensity) and 8 (x, y, water depth, intensity, across track (y), along track (x), heave, heading).*

File set	Number of data streams	Number of pods	Upload time [min]	Conversion time [min]
A	4	1	16.1	147.2
B	4	1	18.3	135.7
C	4	1	19.5	189.1
A	8	1	16.1	177.1
B	8	1	18.3	156.7
C	8	1	19.5	209.0
A	4	8	16.1	95.5
B	4	8	18.3	41.8
C	4	8	19.5	36.5
A	8	8	16.1	146.6
B	8	8	18.3	45.4
C	8	8	19.5	46.2

7. MARISPACE-X: POTENTIAL APPLICATIONS

In the second stage of the Marispace-X project, applications for the four use cases should be tested. Here, we illustrate how these exemplary applications could look like for two of the use cases, Offshore Wind and Munitions in the Sea. They are discussed in a simplified manner, neglecting some of the Intermediaries (***Figure 1***), to illustrate some of the main functionalities.

7.1 Offshore wind

In an offshore wind project, many stakeholders are involved and hence efficient data management, data sharing and contracting are important tasks. We only illustrate a simplified example to demonstrate the underlying concept for these tasks (***Figure 6***).

The figure shows the involved participants, the necessary and authorized data sharing between the participants and the corresponding contracting according to the data sharing. The relations between the participants could be interchanged depending on the specific project needs. The two

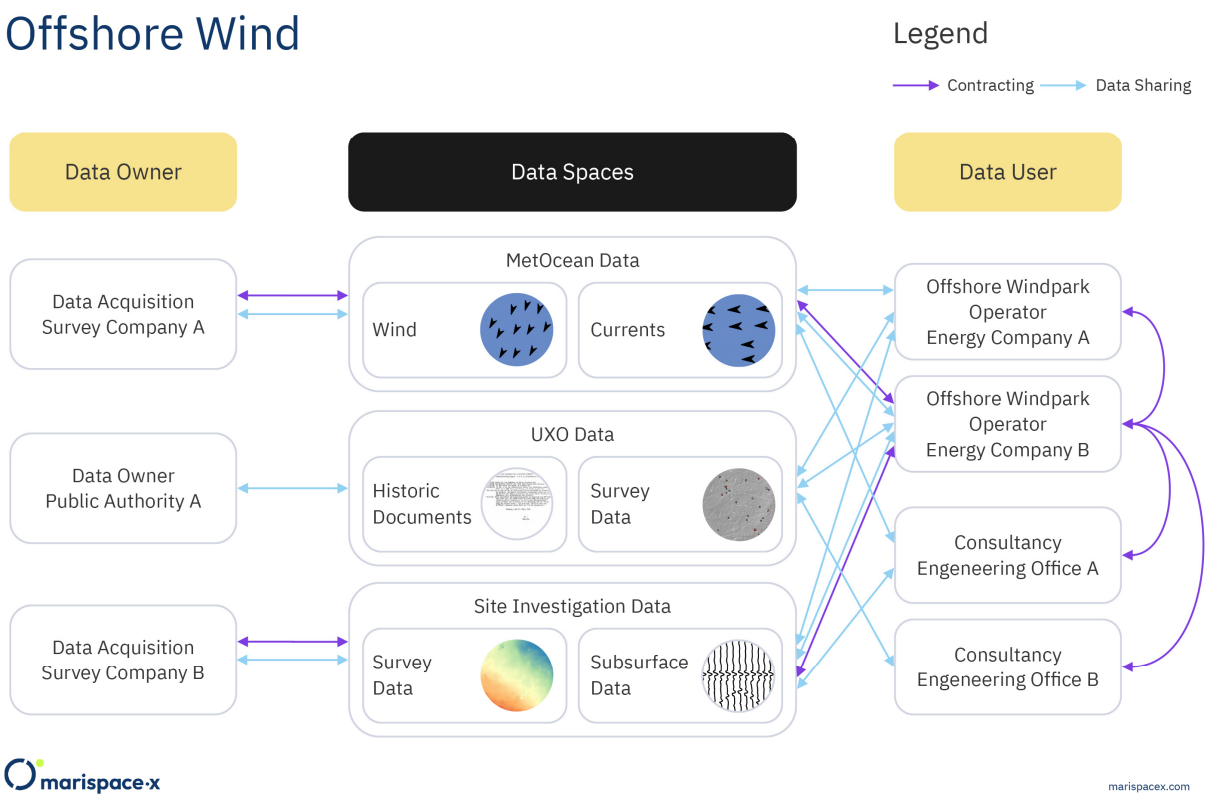

Figure 6. *Sketch of exemplary Offshore wind use case considering data sharing and contracting between stakeholders. The example considers different data (in data spaces) provided by different Data Owners. Different Data Users are only allowed to retrieve specific data from the data spaces as agreed through the contracting. The small exemplary images in the figure are provided by north.io and are adapted from Frey et al. (2021).*

energy companies A and B are planning to build an offshore wind farm and sign the respective contract (***Figure 6***, purple arrow). In this example, they are listed as Data Users, although they often legally own the data. For the planning and site investigations different subcontractors are employed by energy company B (***Figure 6***, purple arrows via data spaces), here denoted as the two survey companies A and B. The subcontractors acquire, analyze, and provide data and hence are technically listed as Data Owners. The results from the survey companies are double checked by an engineering office A that is commissioned by energy company B (***Figure 6***, purple arrow). An UXO (UneXploded Ordnance) desk study is conducted by an engineering office B, also commissioned by energy company B (***Figure 6***, purple arrow). This engineering office retrieves open-access data from the public authority A for its historic research of potential UXO in the construction area. The data can be retrieved from the storage of the respective dataset (data space). Hence, sharing of data is feasible without sending or duplicating the data to several storage devices. Having only one version of each data result available reduces the confusion about duplicated data results stored at different locations. The data spaces are only made available to the respective participants who are authorized.

7.2 Munitions in the sea

There could be different use cases related to munitions in the sea. In this example, a public authority wants to get an overview of potential munition objects in a defined coastal area. Therefore, the public authority commissions an engineering office. The example illustrates the tasks conducted by the engineering office (the Data User) based on the Marispace-X concept (***Figure 7***).

The engineering office, as the Data User, gathers the data within the area of interest defined by the public authority. Via a Broker Service Provider, the Data User could find all available data.

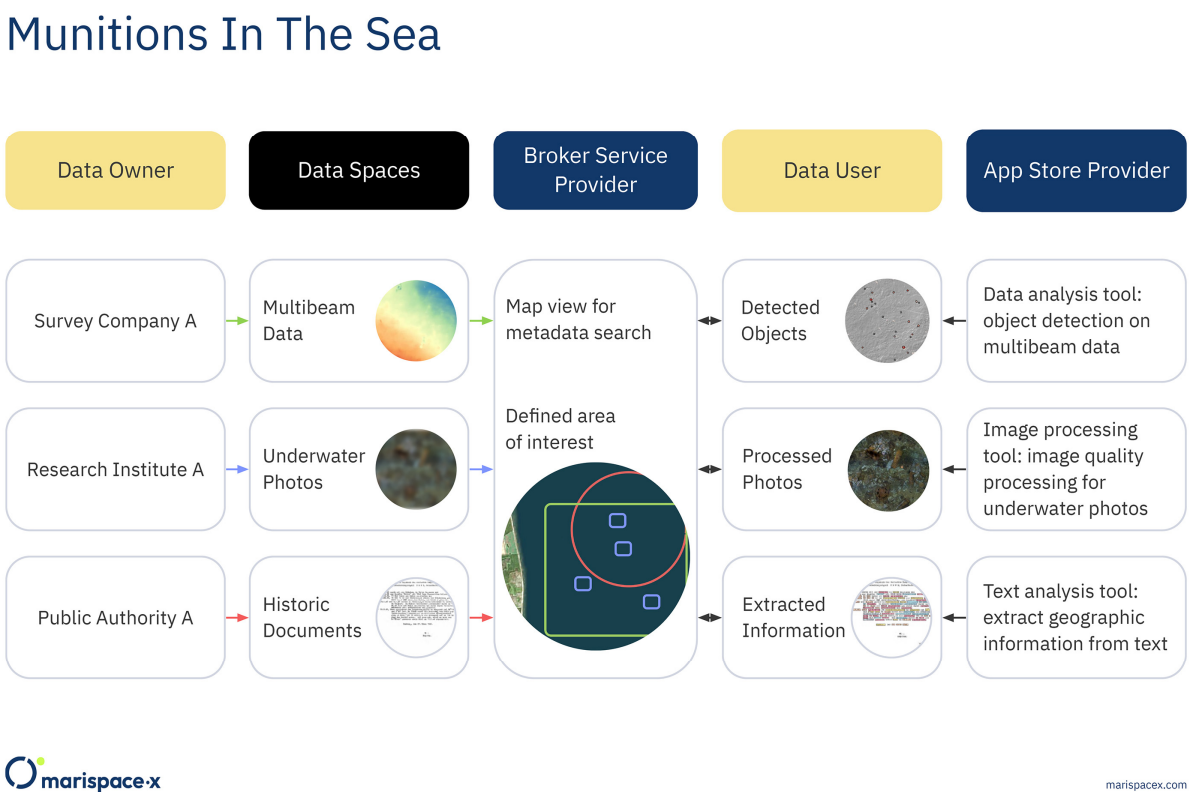

Figure 7. *Sketch of exemplary Munitions in the sea use case where potential munitions objects in a defined area of interest should be shown on a map. The example considers data (in data spaces) provided by different Data Owners. The Data User can find and retrieve the available data via the Broker Service Provider. To analyze the data, the Data User applies different federated services that are retrieved via the App Store Provider. The small exemplary images in the figure are provided by north.io and are adapted from Frey et al. (2021) and Kampmeier et al. (2021).*

The usage options of the data are agreed with the Data Owners through contracting via the Clearing House. Although not shown in the figure, the Clearing House operates at the same step as the Broker Service Provider. The engineering office has its own infrastructure (cluster) available but requires specific tools to analyze the data. The tools are then retrieved as federated services through the App Store Provider and are only used for this specific project. The services used by the engineering office are object detection algorithms, image processing and text analysis tools (***Figure 7***). From the analysis the engineering office can then create a map with all potential munition objects within the defined area, which is then passed to the public authority.

8. CONCLUSION AND IMPLICATIONS

The general concept of a digital ecosystem in Marispace-X according to the Gaia-X standards has been described. These standards are based on European values, which are openness, transparency, sovereignty, and interoperability. It should be noted that the technical standards described here are still in development during the project and hence might slightly change in the future as mentioned in the description of the infrastructure, federated services, and data spaces. Following the general description of the digital ecosystem, potential applications for hydrographic survey data are discussed. This includes the usage of an open-file, cloud-based, and generic data format, namely the *Parquet* format, and a computation example for the upload and conversion of binary data files from a multibeam survey to the generic *Parquet* format using a federated service. In addition, an outlook on potential applications for the use cases in Marispace-X that could be tested during a later project stage are discussed. In general, Marispace-X should allow efficient and safe data management, data sharing, and data processing of heterogeneous maritime data.

The use cases of the Marispace-X project could have future interactions with related projects like Copernicus Marine Services, Copernicus Data and Information Access Services (Copernicus DIAS), the European Marine Observation and Data Network (EMODnet) or the Iliad project (Iliad, 2022). Marispace-X is also part of the endorsed UN Ocean Decade Action 54.2 Interoperability Architecture for a Digital Ocean (United Nations, 2022) within the DITTO programme (DITTO, 2022) which has the goal of developing interoperability architectures for the Digital Twins of the Ocean.

9. ACKNOWLEDGEMENTS

We like to acknowledge the colleagues from north.io for the design of the figures and Matthias Buchhorn-Roth for constructive feedback about the technical components. We acknowledge the German Federal Ministry for Economic Affairs and Climate Action (BMWK) for funding of the Marispace-X project.

10. REFERENCES

AISBL (2022). Gaia-X European Association for Data and Cloud AISBL. Viewed online 29.08.2022: https://gaia-x.eu/who-we-are/association/

Apache Parquet (2022). Apache Parquet Documentation. Viewed online 21.06.2022: https://parquet.apache.org/docs/

Böttcher, C., Knobloch, T., Rühl, N.-P., Sternheim, J., Wichert, U., and Wöhler, J. (2011). Munitionsbelastung der deutschen Meeresgewässer – Bestandsaufnahme und Empfehlungen. Report. Bund-Länder Meeresprogramm. Viewed online 13.09.2022: https://www.schleswig-holstein.de/uxo/DE/Berichte/PDF/Berichte/aa_blmp_langbericht.html?nn=2ab9a364-8dc1-47cb-9e8e-a38d35ca0fef

Calder, B. R., and Masetti, G. (2015). Huddler: a multi-language compiler for automatically generated format-specific data drivers, US Hydrographic Conference (US HYDRO). https://www.researchgate.net/publication/277302751_HUDDLER_a_multi-language_compiler_for_automatically_generated_format-specific_data_drivers

De la Torre, J. (2022). Introducing GeoParquet: Towards geospatial compatibility between Data Clouds, blog article. Viewed online 21.06.2022: https://carto.com/blog/introducing-geoparquet-geospatial-compatibility/

DITTO (2022). DITTO – Digital Twins of the Ocean. DITTO a Global Program of the UN Decade of Ocean Science for Sustainable Development (2021-2030). Viewed online 12.07.2022: https://ditto-oceandecade.org/

Ferguson, J. S., and Chayes, D. A. (2009). Use of a Generic sensor format to store multibeam data, Marine Geodesy, Vol. 18, Issue 4, 299-315. https://doi.org/10.1080/15210609509379762

Frey, T., Kampmeier, M., Seidel, M. (2021). Uncovering the Secrets of German Marine Munitions Dumpsites. High-performance UXO Detection and Visualization. Hydro International 24 (3), 14-17. https://oceanrep.geomar.de/id/eprint/54142/

Gaia-X (2022). Gaia-X – Architecture Document, 22.04 Release. Viewed online 21.06.2022: https://gaia-x.eu/wp-content/uploads/2022/06/Gaia-x-Architecture-Document-22.04-Release.pdf

Heimbigner, D., and McLeod, D. (1985). A Federated Architecture for Information Management, ACM Transactions on Office Information Systems, Vol. 3, Issue 3, 253-278. https://dl.acm.org/doi/pdf/10.1145/4229.4233

Held, P., and Schneider von Deimling, J (2019). New Feature Classes for Acoustic Habitat Mapping – A Multibeam Echosounder Point Cloud Analysis for Mapping Submerged Aquatic Vegetation (SAV), Geosciences, Vol. 9, Issue 5, 235. https://doi.org/10.3390/geosciences9050235

IDSA (2019). Reference Architecture Model 3.0, Version 3.0. Viewed online 21.06.2022: https://internationaldataspaces.org/wp-content/uploads/IDS-Reference-Architecture-Model-3.0-2019.pdf

IDSA (2021). GAIA-X and IDS, Position Paper, Version 1.0. Viewed online 21.06.2022: https://internationaldataspaces.org//wp-content/uploads/dlm_uploads/IDSA-Position-Paper-GAIA-X-and-IDS.pdf

IDSA (2022). IDS RAM 4.0. Viewed online 07.07.2022: https://github.com/International-Data-Spaces-Association/IDS-RAM_4_0

Iliad (2022). Iliad- Digital Twin of the Ocean. European Commission's Horizon 2020 Research and Innovation programme. Viewed online 11.07.2022: https://www.ocean-twin.eu/

Kampmeier, M., Michaelis, P., Wehner, D., Frey, T., Seidel, M., Wendt, J., and Greinert, J. (2021). Workflow towards autonomous and semi-automized UXO Survey and Detection. Proceedings of Meetings on Acoustics 44 (1), 1-11. DOI: 10.1121/2.0001492. https://oceanrep.geomar.de/id/eprint/54425/1/2.0001492.pdf

Kongsberg (2018). Instruction manual – EM Series Multibeam echo sounders Datagram formats, 850-160692/W. Viewed online 21.06.2022: https://www.kongsberg.com/globalassets/maritime/km-products/product-documents/160692_em_datagram_formats.pdf

Leidos (2019). Generic Sensor Format Specification, version 03.09. Viewed online 21.06.2022: https://www3.mbari.org/data/mbsystem/formatdoc/GSF/gsf_spec_03.09.pdf

Lewis, J., and Fowler, M. (2014). Microservices – a definition of this new architectural term, blog article. Viewed online 21.06.2022: https://martinfowler.com/articles/microservices.html

Mell, P., and Grance, T. (2011). The NIST Definition of Cloud Computing, Recommendations of the National Institute of Standards and Technology, Special Publication 800-145, National Institute of Standards and Technology, U.S. Department of Commerce. https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

RFC (2005). Common Format and MIME Type for Comma-Separated Values (CSV) Files, RFC 4180. Viewed online 21.06.2022: https://www.rfc-editor.org/info/rfc4180

SEG (2017). SEG-Y_r2.0: SEG-Y revision 2.0 Data Exchange format, SEG Technical Standards Committee. Viewed online 21.06.2022: https://seg.org/Portals/0/SEG/News%20and%20Resources/Technical%20Standards/seg_y_rev2_0-mar2017.pdf

Teledyne Reson (2019). Data Format Definition Document, 7k Data Format, Version 3.10. Viewed online 21.06.2022: http://www3.mbari.org/products/mbsystem/formatdoc/Teledyne7k/7k_DFD_3.10_package/DFD_7k_Version_3.10.pdf

Triton Imaging (2016). eXtended Triton Format (XTF), Rev. 41. Viewed online 21.06.2022: https://ge0mlib.com/papers/File_Formats/Xtf_rev41.pdf

Totzler, N., and Tschabuschnig, G. (2022). A User Journey to Data Spaces. White paper. Viewed online 11.07.2022: https://github.com/eclipse-dataspaceconnector/Publications/blob/main/White%20paper/A%20User%20Journey%20to%20Dataspaces.md

United Nations (2022). Ocean Decade unveils new set of endorsed Actions on all continents. IOC-UNESCO - United Nations Decade of Ocean Science for Sustainable Development (2021-2030). Viewed online 13.07.2022: https://www.oceandecade.org/news/ocean-decade-unveils-new-set-of-endorsed-actions-on-all-continents/

Wright, D. B., and Wright, C. E. (2017). A Cloud based Solution in Hydrographic Data Processing: The Shift to a Web Centric Software Platform, US Hydrographic Conference (US HYDRO). https://www.researchgate.net/publication/320234486_A_Cloud_based_Solution_in_Hydrographic_Data_Processing

11. AUTHOR BIOGRAPHIES

Daniel Wehner received the B.Sc. degree in geoscience in 2012 and the M.Sc. in geophysics in 2015 from the Christian-Albrechts University (CAU) of Kiel, Germany. He received a PhD in geophysics in 2019 from the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway, where his research was focused on marine seismic acquisition and underwater acoustics. He is currently working at north.io on marine geophysical data quality control, data analysis and interpretation with a focus on munitions in the sea. His research interests are geophysical data acquisition and monitoring techniques. Moreover, he is interested in new sensor technologies. Email: dwehner@north.io

Sergius Dell received his diploma (2009) and a Ph.D. (2012) in geophysics from the University of Hamburg. In 2012-2015, he worked at Fugro and CGG (United Kingdom). Afterwards, he had been researching as postdoc at the University of Hamburg, before he joined TrueOcean GmbH in 2020. His main focus is processing of geospatial data in the cloud. Email: sdell@trueocean.io

Adrian J. Neumann holds a B.A. Hons in European Business from Berlin School of Economics (2005) and a Master of Science in Strategic Marketing from Cranfield University UK (2006). After several years in the private sector, Adrian joined the Institute for Security Policy's Center for Maritime Strategy & Security at Kiel University (ISPK), Germany in 2014, focusing on regional maritime security challenges and societal resilience research. Since 2020 Adrian works at the software developer north.io where he helps to take on the global challenge of sea-dumped munitions through digitalization, and where he co-coordinates the Marispace-X project. Email: aneumann@north.io

Jann Wendt received his B.Sc. in environmental geography and management in 2014 from the Christian-Albrechts-University (CAU) of Kiel, Germany. He is CEO and founder of several German based tech companies (north.io GmbH, TrueOcean GmbH and NatureConnect GmbH) that combine digitalization, environmental protection and the fight against climate change. His work is focusing on Spatial Big Data, Cloud Technologies and AI in the domain of environmental protection. Email: jwendt@north.io