AGILE: GIScience Series, 6, 38, 2025. https://doi.org/10.5194/agile-giss-6-38-2025
Proceedings of the 28th AGILE Conference on Geographic Information Science, 10–13 June 2025.
Eds.: Auriol Degbelo, Serena Coetzee, Carsten Keßler, Monika Sester, Sabine Timpf, Lars Bernard
This contribution underwent peer review based on a full paper submission.

Making Geospatial Data Assets Discoverable and Accessible in the Green Deal Data Space

Joan Maso[1] Alba Brobia[1], Raul Palma[2], and Ignacio Elicegui[3]

[1] Grumets Research Group, Centre de Recerca Ecològica i Aplicacions Forestals (CREAF), Cerdanyola del Vallès, 08193 Bellaterra, Spain;
[2] Poznan Supercomputing and Networking Center; Wieniawskiego 17/19, 61-704 Poznań, Poland
[3] ATOS, C Albarracín, 25, 28037 Madrid

Correspondence: Joan Masó, joan.maso@ieee.org

Abstract. A Data Space is a distributed system defined by a governance framework that enables secure and trustworthy data transactions between participants. The architecture of a data space is defined by the data space protocol and composed by several building blocks. Metadata catalogues are a fundamental building block that provides discoverability and the Data Space Connectors provide accessibility. This paper proposes an approach for the combination of geospatial standards for data discovery and the data space protocol and describes a practical implementation in the context of the AD4GD project. The presented architecture provides access to internal users via a SaaS cloud based on MinIO for data storage, GeoNetwork for data discovery, and Docker for data processing. This infrastructure constitutes an experimental data space node provided by the All Data for the Green Deal (AD4GD) project. The AD4GD project is conducting research and development for a future Green Deal Data Space.

Submission Type. Short paper

BoK Concepts. [WB4] Resource Discovery, [WB7] Web Application development elements

Keywords. Data space, metadata, FAIR, discoverability

1 Introduction

Data is considered a vital part of the digital economy promoted by the European Commission under the European strategy for Data (EC, 2020). The question of how to organize big data to help decision-makers to deliver better fact-based decisions remains (Tan et al., 2017). The main factor for enabling the digital economy is the capacity to exchange data. However, several barriers exist including a lack of interoperability amongst data exchange (e.g. non-machine-readable datasets), missing or incompatible metadata, a lack of trust, fear of losing control of shared data (Jahnke et al., 2024), unknown data quality, data out of date, the lack of clear and adequate license and limited interest in exposing the data (Conde et al., 2024). To remediate some of these aspects, the concept of a data space is proposed.

A data space is a distributed system defined by a governance framework that enables secure and trustworthy data transactions between participants (Poikola, et al., 2023). Data spaces provide flexible correlation of disparate data sources, allowing the spontaneous discovery and exploration of relationships, patterns and trends that may not be apparent in isolation (Theissen-Lipp, J. et al., 2024). In a data space, transactions are mainly about assets a.k.a. datasets and services dealing with data.

In order to make assets known to the actors in the data space, and eventually to the rest of the world, data producers need to generate descriptions of their offerings and create and maintain a mechanism to disseminate their existence. This may sound new for the data space community (Labadie, C. et al. 2020) but the Spatial Data Infrastructures (SDIs) have been building data portals for more than 25 years (Guptill S.C., 1999) based on metadata catalogues. A data catalogue (DC) is a fundamental part of data infrastructures as it enables the inventory and locating of available data in a data ecosystem. Data portals and Data Space Catalogues are two of the seven typologies of data catalogues identified

by Jahnke and Otto (2023). The main difference between the data portal and the data space approach is that data portals have a unidirectional relationship where only the provider shares data to the consumer, while data spaces propose a relationship with data catalogues contributed both by producers and consumers and allowing for a bidirectional data exchange. Jahnke and Otto also define the marketplace as an evolution of the data space with additional modules for monetization of data transactions.

From a technical point of view, all typologies of catalogues require descriptions of their assets that should not be provided in plain text but structured in a set of elements that can be easily stored in fields of a database. These descriptions are commonly known as *metadata*. In 2003 the first ISO standards for geospatial metadata was released based on previous efforts mainly for the Federal Geographic Data Committee (ISO 19115). For each asset, a metadata record describing it is produced and stored in a catalogue. In 2007 the Open Geospatial Consortium (OGC) released the OpenGIS Catalogue Services Specification (CSW); a standard web service that allows for query of geospatial metadata.

Data Catalog Vocabulary (DCAT) defines a common schema for describing datasets and services, including properties such as title, description, keywords, and access information. DCAT is provided as a Resource Description Framework (RDF) vocabulary which standardizes the descriptions for datasets, services, and catalogues, making it easier to discover, share, and reuse data resources across different platforms and organizations. This vocabulary was standardized in 2014 by the W3C Government Linked Data (GLD) Working Group. GeoDCAT-AP is an extension of DCAT-AP (DCAT Application Profile) specifically designed for describing geospatial datasets, dataset series, and services. It provides an RDF syntax binding for metadata elements defined in the core profile of ISO 19115:2003 and those defined in the framework of the INSPIRE Directive (De Cock, J. et al. 2024).

The data space protocol (most recent version 2024-1) is a set of specifications initiated by the International Data Spaces Association (IDSA) for data sharing between entities in a data space. For the publication, management and discovery of metadata, the data space protocol defines a catalogue protocol that describes how descriptions of assets are deployed inside catalogues represented using the DCAT:Catalog class and its properties and how access and usage control of data assets are expressed as Open Digital Rights Language (ODRL) policies. From this perspective, the catalogue is a collection of assets published in the form of DCAT datasets and services records. A catalogue must include at least one data service that references the service (a.k.a. a connector) that provides these datasets. It is via the connectors (software components for enabling trust and sovereignty in data sharing) that assets can be managed and contracts between consumers (users) and data providers can be stablished and negotiated according to a set of rules in a data space.

The objective of the presented research is to demonstrate how metadata can be transported among the different components forming part of a node of a data space. While work has been done on how to deal with metadata in a data space, no previous research on how to transport and transform metadata between geospatial technologies, cloud technologies and data spaces technologies has been found, and this fact justifies the novelty of the approach.

In its role of supporting the European Green Deal policies aimed at combating climate change, the Green Deal Data Space deals mainly with environmental data which is intrinsically geospatial. Environmental data produced by the government and public administrations in Europe shall follow the INSPIRE directive that recommended the use of ISO19115 metadata. On the other hand, the Green Deal Data Space should follow the IDSA recommendations and adopt DCAT. This paper describes a proposal on how to connect ISO19115 metadata catalogues into the connector DCAT based metadata catalogues.

2 Materials and methods

2.1 Data and metadata cataloguing and accessing

Due to most of the data necessary to run the use cases of the AD4GD is of geospatial nature, the project adopted the GeoNetwork catalogue system (v4) that offers native support for ISO19115/ISO19139 geospatial metadata. It provides a GUI for discovering and publishing descriptions of datasets, allowing organizations to centralize geospatial metadata from different sources and make them accessible through standardized web services such as the OGC CSW.

Even if GeoNetwork allows linking data as attachments to the metadata, it is not designed for direct data storage. Thus, AD4GD adopted an IaaS cloud to store and process vector data files called MinIO. MinIO allows to store vast amounts of data using ML techniques and AI cloud storage. MinIO has an API compatible with the Amazon S3 cloud storage service. Several AI processing algorithms are also setup in the same cloud infrastructure, so they can be trained with the data stored in MinIO. There are two exceptions: the constant sensor data flows are stored in a OGC SensorThings API and the gridded base data (raster data) are stored and organized in an Open Data Cube. In both exceptions and

when needed, queries to a subset of the data can be formulated, and the resulting snapshots stored in the MinIO for convenience. In addition, the Eclipse Data Connectors (EDC) are also deployed in the same cloud that exposes some of the data stored in MinIO.

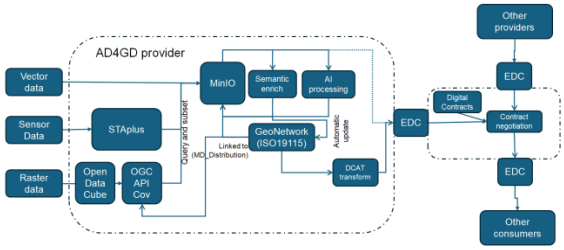

Figure 1. AD4GD General Architecture.

2.2 Client application

To illustrate how to interact with the different components of the architecture we have developed a client application that is called TAPIS (Tables from APIs). TAPIS is an API explorer and a table manager. TAPIS reads data and metadata from some supported APIs and some data file formats and structures the data as tables that can be managed and transformed. Internally, everything is a table that has columns that represent fields and rows that represent records. Once the data is imported, they can be directly viewed, edited, semantically enriched, filtered, joint, grouped by, aggregated, etc. as well as presented as bar charts, pie charts, scatter plots, and maps.

2.3 Data and Software Availability

The developed components in the AD4GD research, including the full code of TAPIS and other code sniped used are published under a permissive MIT license and available at https://github.com/ad4gd. The access to the AD4GD cloud infrastructure cannot be granted due to operational costs. The AD4GD data space access can only be granted under demand to the members of the experimental GDDS due to the nature of the secured infrastructure.

3 Results

3.1 Metadata production

Every participant in the use cases of the AD4GD project had access to the GeoNetwork system to create and maintain the metadata describing the objects in MinIO. The creation of good metadata is time consuming and sometimes there are more than one way of documenting the same property. To ensure some degree of harmonization, a set of good practices were defined and a person into the project was assigned to review the metadata records and ensure that the good practices are correctly applied. In here we describe four good practices that were considered particularly relevant.

- **Start from a template**

There are several fields that need to be populated to create a good metadata record. It is important to start documenting metadata following a template. In our case we have created a template for object based data and another for the gridded data cube based data.

- **Link to vocabularies**

One challenge in AD4GD was to find semantic interoperability solutions for integrating of heterogenous data in the GDDS. One manner is starting by encoding and linking data concepts to ISO19115 MD_Keywords class. The ISO19139 XML schema facilitates the possibility to transform any string into a link. This can be done by replacing <gco:CharacterString> element by a <gmx:anchor>. Via this mechanism, it is possible to link individual keywords associated with variables (e.g. temperature, precipitation) to their definitions in the context of a related thesaurus or controlled vocabulary. In the same way, in CI_Citation class it is also possible to associate a URI to a Citation of a thesaurus as a "collection of concepts" using a link in the Title field. Although the revised ISO19115-1 already have a URI in Citation, it is absent in the original ISO19115:2003. In AD4GD we opted to use ISO 19115:2003 as it remains the most widely used version of the standard, and it is still the one recommended in the INSPIRE implementing rules.

- **Data quality and provenance**

QualityML is a profile of the ISO (e.g. ISO19115 and ISO19157) providing a set of rules for precisely documenting quality information and controlled vocabularies for concepts in a linked data style. The vocabulary encompasses quality indicators, quality measures, quality domains and matrices. Quality matrices use statistic expressions for describing uncertainties, whenever possible, from the UncertML dictionary (Reichert, P. et al. 2010). In addition, the QualityML documentation (Ninyerola, M et al. 2014) provides a set of examples on how to encode data quality in ISO19115/19139 (with and without extending the ISO metadata) that are used as a reference in AD4GD.

- **Link to the actual data**

In addition to discovering resources, the purpose of a metadata catalogue is to facilitate the access to the data described in it. GeoNetwork allows to store any data file as attachment to the metadata records or to associate a URL to the data is stored MinIo that then can be linked

and accessed from external services according to the ISO19115:2003 MD_Distribution class. This class provides the possibility to describe online access methods for the digital distribution of a dataset (MD_DigitalTransferOptions), identification of the distributor (MD_Distributor) and the format of the distribution (MD_Format).

3.2 AD4GD Internal Access

MinIO is a storage facility compatible with Amazon S3 that uses the same style of file folder structure that the common desktop operating systems use (that is also adopted by the personal mass market cloud SaaS such use Google drive, Microsoft Onedrive). While discovering data reading file names in a folder classification works well for a small amount of files, it has two disadvantages: it does not provide rich descriptions of the files discovered and it becomes impractical when many files belong to more than one folder. Unfortunately, MinIO is poor in terms of metadata and it is limited to a number of headers starting by "X-Amz-Meta-". This allows for a minimum set of metadata terms to be included in MinIO objects using dcat keys.

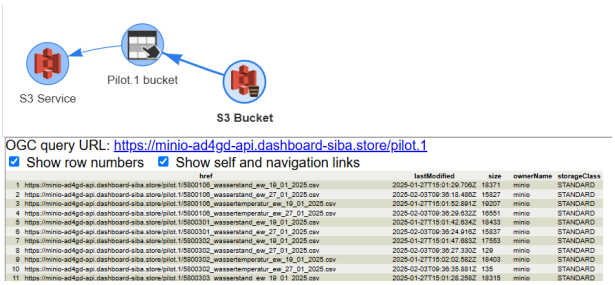

Figure 2. AD4GD MinIO bucket represented in TAPIS.

Due to the lack of metadata in MinIO, for each resource in AD4GD MinIO instance, a metadata representation in GeoNetwork was created. The GeoNetwork is the way to provide discoverability in the AD4GD internal system. Is is also part of the accessibility because a full metadata record includes a MD_Distribution pointing to the MinIO object (the data). The MinIO object has a metadata term that links back to the GeoNetwork record.

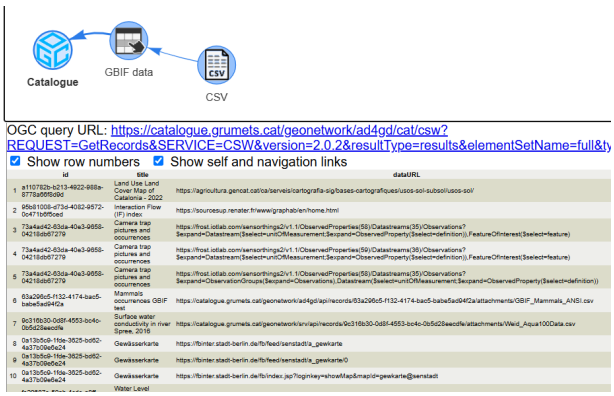

Figure 3. AD4GD GeoNetwork represented in TAPIS as well as the selection of a CSV file form the catalogue.

3.3 GDDS External Access

The final components that will be used in the GDDS is still to be determined but in AD4GD we have selected the Eclipse Data Connector as it is one of the mature technologies available. Under this assumption, each node in the GDDS will be accessed by a connector and AD4GD is presented by a connector deployed by the partner ATOS. This data connector is connected to the MinIO object storage and exposes a selection of the objects of the AD4GD project. To simulate the access from another node in the GDDS, another node is deployed by the partner PSNC. To facilitate the discovery in the GDDS, a federated catalog is deployed exposing metadata about all assets accessible thought the data space. PSNC has deployed a federated catalog to do the simulation. This catalogue exposes some DCAT keys as well as providing a temporary id of the assets for each asset. The temporary ids are used to negotiate a contract and once this contract is approved, request for the data transfer is issued. In our implementation, the metadata in the EDC federated catalog is encoded in DCAT and based on the GeoNetwork original metadata records.

The following figures illustrate the process done by a client that uses the PSNC connector to negotiate a contract to the AD4GD connector asset and how the data is transferred.

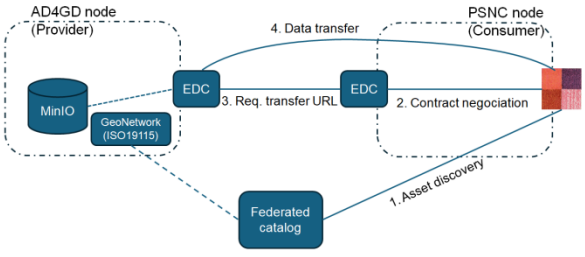

Figure 4. Discovering and Accessing of an asset in the GDDS.

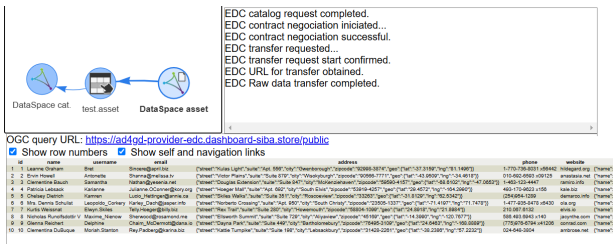

Figure 5. A test assed is discovered in the federated catalogue, negotiated and transferred to the client through the data connector protocol in TAPIS. This practical example corresponds to the previous conceptual diagram.

3.4 Lakes in Berlin data workflow

The water use case in AD4GD aim to improve the health of small lakes in the city of Berlin, Germany, by facilitating access to data about their status. The use case overcomes the current data gaps by combining information from different sources such as measurements done by water sampling, IoT sensors, remote sensing satellites and citizen science programs that are regularly updated. The main datasets are about water level, water temperature and weather, and they are made available every day. Once a new dataset emerges, it is saved in the AD4GD MinIO object storage and a metadata record describing it is included in the GeoNetwork catalogue. This triggers an event that calls the semantic enrichment process developed by PSNC. The enrichment process produces a JSON-LD version of the same data, including links to external recognized vocabularies such as those provided by the European Environmental Agency (https://www.eea.europa.eu/help/glossary/eea-glossary) or the generic QUDT (https://www.qudt.org/). This new JSON-LD version is also stored in MinIO, but it does not include and metadata inside. That is why, it documented as another distribution method in the original metadata record. Then, semantically enriched datasets are exposed in the data space via the AD4GD EDC. This inclusion results in a new asset available for the GDDS and the ISO19115 core metadata exposed in the EDC federated catalogue.

4 Conclusion and future work

The AD4GD project successfully adopted the Data Spaces Support Centre (DSSC) building blocks (DSSC Blueprint, Version 1.0, 2024) and some OGC Web Services and APIs to establish the elements required to support the implementation of the GDDS. The project gives support to external users via a Publication and Discovery building block, materialized as the EDC that is integrated in the GDDS federated catalogue.

In a data space, data is exchanged between data connectors. For demonstration purposes a PSNC node was added to the data space. A client application connects to PSNC API, negotiates a contract and gets authorization and an endpoint for data transfer.

Due to the proliferation of different technologies specified by different organizations (e.g, IDSA, OGC etc), the AD4GD system has three copies of some elements of the metadata for each dataset. We consider the GeoNetwork the master copy and the most comprehensive and the one is used to propagate the metadata to the minimum set in MinIO and the DCAT metadata in the EDC assets. The main drawback of this approach is the limitations on the mapping between metadata standards and the need for the participants in the data space to have access to an EDC instance to get the data.

The next step is to make the metadata management more automatic and include data provenance documentation in the processing software deployed in the AD4GD node. Another aspect that needs to be improved is the guidelines for deciding what datasets should be exposed in the data space, by establishing a protocol for validating the data and metadata and associating a policy of access conditions.

There is also a need to combine open data with protected data in the data space. For the moment, we expose open data by associating empty policies to assets and allowing for a trivial negotiation of void contracts, but we are still imposing the need for having access to an EDC connected to the data space to be able to discover and get the data.

We are also considering how the concept of data space can be extended to provide data services and processing time in the data space.

Declaration of Generative AI in writing

The authors declare that they have not used Generative AI tools in the preparation of this manuscript. All intellectual and creative work, including the analysis and interpretation of data, is original and has been conducted by the authors without AI assistance.

Acknowledgements.

This research is funding by the European Union under the Horizon Europe project AD4GD (No 1010610001) and more4nature (No 101133983).

References

Conde, J., Pozo, A., Munoz-Arcentales, A., Choque, J. and Alonso, Á., Fostering the integration of European Open Data into Data Spaces through High-Quality Metadata, arXiv preprint arXiv:2402.06693, 2024

De Cock, J., Escriu, J., Fragkou, P., Perego, A., Schiltz, A. and Van Nuffelen, B. GeoDCAT-AP 3.0.0, 04 October 2024. Web: https://semiceu.github.io/GeoDCAT-AP/releases/3.0.0/

European Commission. *A European Strategy for Data*. Publications Office of the European Union, 2020. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066

European Parliament and Council of the European Union**.** Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Official Journal of the European Union, vol. L108, pp. 1–14, 2007. Web: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32007L0002

Guptill, S. C.: Metadata and data catalogues. Geographical information systems, 2, 677-692, 1999.

International Organization for Standardization. ISO 19115:2003 – Geographic Information: Metadata. Geneva, Switzerland: ISO, 2003. Web: https://www.iso.org/standard/26020.html

Jahnke, N., and Otto B.: Data catalogs in the enterprise: applications and integration, Datenbank-Spektrum 23, no. 2, 89-96, 2023.

Jahnke, N., Spiekermann M., and Möller F., Federated Data Catalogs for Data Sharing–Towards Design Principles, Thirty-Second European Conference on Information Systems (ECIS 2024), Paphos, Cyprus, 2024.

Labadie, C., Legner, C., Eurich, M. and Fadler, M., 2020, June. FAIR enough? Enhancing the usage of enterprise data with data catalogs, IEEE 22nd Conference on Business Informatics (CBI) 1, pp. 201-210), 2020

Ninyerola, M., Sevillano, E., Serral, I., Pons, X., Zabala, A., Bastin, L., and Masó, J. QualityML: A dictionary for quality metadata encoding. In EGU General Assembly Conference Abstracts, pp. 10452., 2014.

Poikola A., Verdonck B., Joosten R. Guggenberger T. and Salminen S.: DSSC Glossary, Data Spaces Support Centre (DSSC), available at: https://dssc.eu/space/Glossary, last access: 27 December 2024, 2023.

Reichert, P., Schirmer, M., Schuwirth, N., & Beven, K. (2010). Uncertainty Markup Language (UncertML). *Environmental Modelling & Software, 25*(2), 155-160. https://doi.org/10.1016/j.envsoft.2009.06.009

Tan, K.H., Ji, G., Lim, C.P. and Tseng, M.L., Using big data to make better decisions in the digital economy, International Journal of Production Research, 55(17), pp.4998-5000, 2017

Theissen-Lipp, J., Hemid, A., Lange, C. and Gillmann, C., Towards a Federated Data and Service Catalogue, The 3rd International Workshop on Semantic Interoperability in Data Spaces, October 01, 2024, Budapest, Hungary, 2024.