AD4GD semantic interoperability approach for a Green Deal Data Space

Raul Palma, Bogusz Janiak, Poznan Supercomputing and Networking Center (PSNC),
rpalma@man.poznan.pl, bjaniak@man.poznan.pl
Rob Atkinson, Piotr Zaborowski, Alejandro Villar, Francesca Noardo - Open Geospatial Consortium (OGC)
ratkinson@ogc.org, pzaborowski@ogc.org, avillar@ogc.org, fnoardo@ogc.org
Alba Brobia Ansoleaga,Ivette Serral, Joan Maso Pau (CREAF)
a.brobia@creaf.uab.cat,ivette@creaf.uab.cat, joan.maso@uab.cat
Lucy Bastin (Aston University) l.bastin@aston.ac.uk

One of main goals of AD4GD is to design and implement the building blocks to support a Green Deal Data Space that will provide interoperability support for the heterogeneous and increasingly growing set of data and services available in the various action areas addressed by the Green Deal (GD) (e.g., climate, energy, industry, agriculture and biodiversity) through definition, sharing and assembly of standardized building blocks and that will directly contribute towards a European GD Data Space. Indeed, interoperability is one of the main categories of building blocks identified by the OPEN DEI project[1] The three building blocks related to interoperability include:

- Semantic Harmonization which establishes a common semantic representation of data in data exchange payloads. Combined with the Data Exchange APIs building block, this ensures full interoperability among participants.
- Data Exchange APIs, which facilitate the sharing and exchange of data (i.e., data provision and data consumption/use) between data space participants.
- Data Provenance and Traceability, which provide the means for tracing and tracking in the process of data provision and data consumption/use. It provides the basis for a number of important functions, from identification of the lineage of data to audit-proof logging of transactions. Provenance information will enable the participants to maintain data provenance as part of the metadata during the process of data exchange.

In line with this framework, AD4GD is dealing with the building blocks addressing common interoperability challenges, defining a modular semantic model for GD related data, including the provenance metadata, and APIs, leveraging and reusing existing standards as much as possible, and providing the mechanisms to enable different systems to exchange data with unambiguous meaning, and to enable an integrated data access for the execution of advanced data analytics to support the project pilots.

Interoperability challenges can be addressed at different levels. The International Data Space Association (IDSA), for example, relies on the European Interoperability Framework (EIF) [1], which defines four layers: technical, semantic, organizational and legal. Although AD4GD is interested in all the interoperability layers, in this document we are focusing on the semantic interoperability aspects and the relationships to the technical interoperability layer which in practical terms constrains the common vocabularies. In particular, semantic Interoperability requires describing parts of information systems that need to be exploited by other components in a system. The available mechanisms for describing the semantics of data are varied and evolving and it's not possible to define a single solution that can be used for all cases, however the technical challenges and best practice options to address these can be identified. The level of expertise required and potential variability of semantic description mechanisms means that reusable

[1] Design Principles for Data Spaces – Position Paper - https://design-principles-for-data-spaces.org/

patterns need to be identified and applied consistently. This is too complex to achieve for complex and complete application data models, and requires standardization of specific parts as easy to use "building blocks" for system implementation.

In this sense, our Data Space concept also gets inspiration from the same approach as the AgriDataSpace project[2], which focuses on an incremental model towards a full semantic integration based on the pay-as-you-go (PAYG) data management approach [2]. Inspired by the 5-star rating scheme defined by Tim Berners-Lee[3] that helps data publishers to evaluate how much their datasets conform to the linked data principles, this model provides flexibility by reducing the initial cost and barriers to joining the dataspace. The more the investment made to integrate with the support services, and the standards used to describe aspects of semantics in different services, the better (faster, cheaper, more robust, repeatable and transparent) the integration achievable in the dataspace. The five levels (and stars) of the model include:

1. Basic (minimum): data source is published in the data space with limited or no integration with support services.
2. Machine-readable: the data source is publishing data in a machine-readable format. This enables services to provide a minimal level of support with basic functionality (e.g., browsing the data) where available basic interfaces are exposed.
3. Basic integration: the use of a non-proprietary (data) format, and machine-readable metadata, enable support services to provide essential services at the data-item/entity level with support for simple functionality (e.g., keyword search).
4. Advanced integration: the data is integrated with most support service features (e.g., structured queries) with an awareness of its relationships to other data sources participants with basic support for federation.
5. Full semantic integration: the data is fully integrated into the support services (e.g., question answering) and linked to relevant participants. It plays its full role in the global view of the data space.

The advanced integration and full semantics layers are both covered by the semantic interoperability aspects, where there is an agreement on the use of open standard formats, data models and APIs. Hence, in order to enable different systems to exchange data with unambiguous meaning, and the provision of an integrated view over data from different (and heterogeneous) data sources, there is a need for an agreed common information model, or lingua franca, identifying and defining the data elements relevant to the application domain (i.e., Green Deal in case of AD4GD) along with their associated semantics/meaning for information exchange. Essential Variables[4] (EVs) are considered in AD4GD the basic harmonization piece in terms of semantics, thus becoming the common understanding thesaurus among pilots. The original EVs descriptions already defined by every community have been atomized into fundamental pieces to describe an observation, but also in terms of the products that can be derived from these observations (i.e. recommended spatial/temporal resolutions, metrics to be monitored, etc). In this conceptual separation, I-ADOPT Framework[5] for describing environmental observations has been adopted. With this new structure, a unique thesaurus repository for every EV set has been created in the OGC RAINBOW definition server. With this, unique identifiers for every EV product exist and all datasets in a Data Space can refer to this common dictionary for setting a clear and precise meaning of its attributes. For the moment, a preliminary test has

[2] https://agridataspace-csa.eu/

[3] https://5stardata.info/en/

[4] Anthony Lehmann, Paolo Mazzetti, Mattia Santoro, Stefano Nativi, Joan Masó, Ivette Serral, Daniel Spengler, Aidin Niamir, Pierre Lacroix, Maria Paola Ambrosone, Ian McCallum, Nataliia Kussul, Petros Patias, Denisa Rodila, Nicolas Ray, Grégory Giuliani. Essential earth observation variables for high-level multi-scale indicators and policies. Environmental Science & Policy, Volume 131. 2022. https://doi.org/10.1016/j.envsci.2021.12.024

[5] Magagna, B., Moncoiffé, G., Devaraju, A., Stoica, M., Schindler, S., Pamment, A., & RDA I-ADOPT WG.: InteroperAble Descriptions of Observable Property Terminologies (I-ADOPT) WG Outputs and Recommendations (1.1.0). https://doi.org/10.15497/RDA00071, 2022

been done with Essential Biodiversity Variables defined by GEOBON and EuropaBON. Next, all other EVs will follow the same methodology.

On top of that, it is necessary to provide the mechanisms to transform/lift data into this common model, and to integrate it with other related datasets, providing a harmonised data layer that can be exploited by the different data analytics tools. As mentioned in the EIF, Linked Data (combined with a data-driven design) can substantially improve semantic interoperability. Indeed, Linked Data is nowadays one of the most popular methods for publishing data on the Web due to the benefits it can provide (e.g., improved data accessibility, support for data integration and interoperability, knowledge discovery and linking). Such an approach has been demonstrated (and it is being demonstrated) in different projects (e.g., DEMETER, SIEUSOIL, ILIAD, OPEN IACS, DataBio, etc.), where Linked Data has been used or is being used as a federated layer to support large scale harmonization and integration of a large variety of data collected from various heterogeneous sources. Following a similar approach, AD4GD will rely on the implementation of Linked Data pipelines, which in turn rely on the agreed information model, for the representation of data.

Finally, the AD4GD approach considers that different components or systems will implement common open APIs to expose and consume the data, leveraging existing standards, particularly from OGC, in order to boost the interoperability potential with existing and future components. In particular, the combined use of a common information model and open APIs will facilitate the integration of heterogeneous data sources, such as satellite-based earth observation-, climate model-, IoT- and citizen science data and measurements provided by scientists as well as other data created by administrations such as INSPIRE data, into a common European GD data space.

Green Deal Information Model

The AD4GD Green Deal Information Model (GDIM) is a common vocabulary aiming at providing the basis of a common GD data space and enabling the interoperability of different systems, potentially from different vendors. Based on the GDIM, i) data producers/integrators will be able to adapt and apply existing tools to pre-process, integrate and harmonize data from different sources (see next section); ii) service providers will be able to develop lightweight service wrappers and translators, also known as data providers and consumers, which will enable the different tools/platforms to expose and consume data in an interoperable form.

GDIM has been designed and implemented following a modular approach in a layered architecture. Based on the analysis of the state of the art and the initial analysis of the modeling requirements, GDIM is implemented by reusing and building partially over the Agriculture Information Model (AIM), which in turn reuses and aligns various relevant cross-domain standard ontologies and vocabularies, particularly from OGC and W3C. The same approach has been taken to transpose the AIM in the Ocean domain in the ILIAD project to create the Ocean Information Model (OIM), currently under development. AIM is currently under the process of becoming an OGC standard. Hence, analogously to AIM, defines the following layers (depicted in Figure 1):

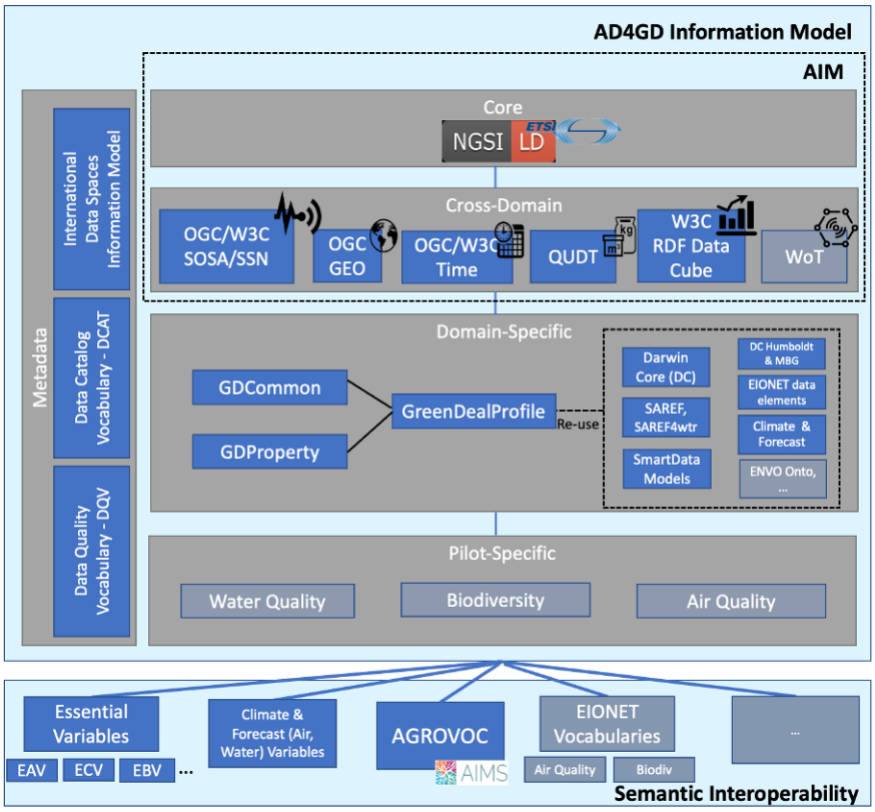

Figure 1. Overview of the layers of the first release of the AD4GD Information Model. Dark blue boxes are implemented, while the grey ones are not (yet). Hence, the pilot-specific modules (if needed) will be specified in the next iteration after pilots have matured.

- the meta-model layer, defining the model building blocks and enabling the back-and-forth conversion between datasets that are based on the property graph model and linked data datasets
- the cross-domain layer, defining relevant concepts and properties that are common across multiple domains, and enabling the interoperability with existing standard models and vocabularies
- the domain layer defining GD related concepts and properties covering different aspects of interest of GD applications, and enabling the integration of relevant vocabularies in the area.
- the pilot-specific layer defining additional concepts and properties that are of specific use for particular applications (if needed)
- a metadata model layer that can be used to describe datasets, services or applications in AD4GD.

GDIM is realized as a suite of OWL ontologies (serialized as Turtle), establishing alignments between the reused standards and well-scoped dominant models to enable their interoperability and the integration of existing data. From these ontologies we generate other related semantic artifacts, including JSON-LD contexts and SHACL shapes (in the future), to facilitate the adoption by service developers/providers, and the validation of data at the semantic level, respectively.

All the AD4GD modules and related resources are publicly available in the AD4GD GitHub repository: https://github.com/AD4GD . All the ontologies generated, as well as the corresponding JSON-LD contexts, use persistent identifiers that are resolvable. The base namespace for OIM is: https://w3id.org/ad4gd/model (which resolves to the AD4GD GD profile module, the main entry point to the IM). The GDIM will be accessible via the OGC Registry for Accessible Identifiers of Names and Basic Ontologies for the Web (RAINBOW) server (formerly OGC Definitions Server), which supports the profiling of complex models to provide a pathway to multiple implementation patterns.

Methods and tools for data harmonisation

AD4GD is implementing a data harmonisation and integration approach based on the adoption of Linked Data as a federated layer, combined with the use of knowledge graph technologies, where data is made available and combined according to common ontologies/vocabularies. This approach has been showcased and demonstrated in different projects, including DEMETER for the agriculture domain, SIEUSOIL for the soil domain, ILIAD for the Ocean domain, OPEN IACS for the CAP framework, and others. The approach carries out different tasks for the generation and publication of Linked Data (as depicted in FIgure 2) in line with best practices and guidelines.

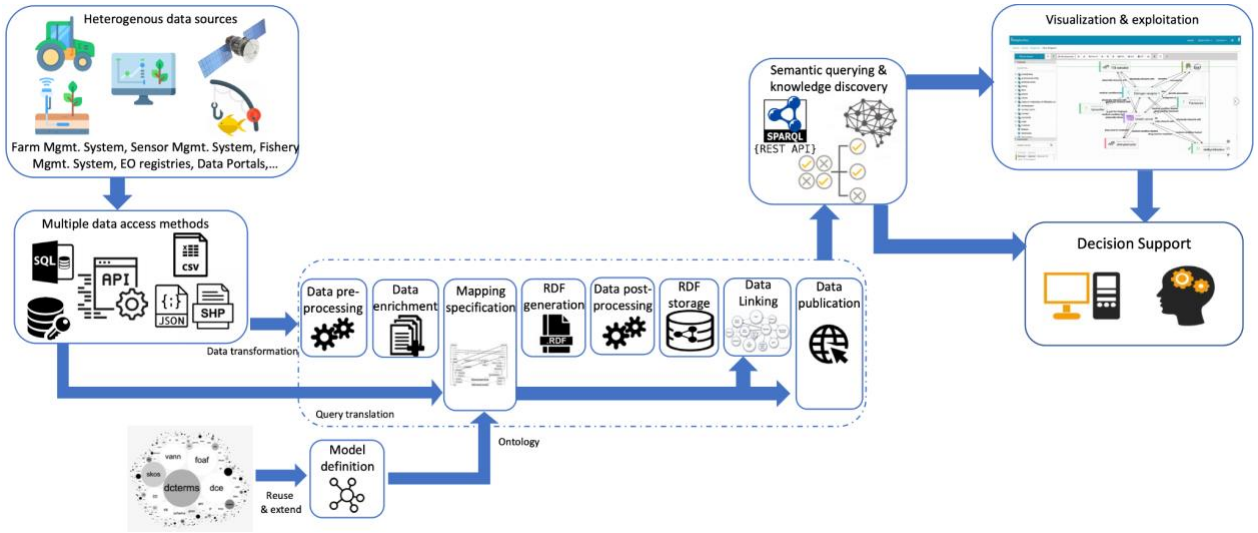

Figure 2. Linked Data pipelines for data harmonization and integration.

As depicted in the figure a key input for the pipelines are the selected target ontologies/vocabularies. In the case of AD4GD this is the Green Deal Information Model (GDIM) previously described. In order to enable the access to the harmonised and integrated data, the AD4GD approach considers different possible interfaces. Data can be accessed directly via semantic queries (i.e., (Geo-)SPARQL), but in order to facilitate its usage by service providers, data providers and other users, the pipelines consider the possibility to generate Restful API that can be created on-the-fly from SPARQL queries, and the (semi-)automatic generation of standard APIs, mainly from OGC, such as the SensorThings API or Features API.

AD4GD brings various data harmonization tools. These include:

- Data Preparation and Integration Pipelines (DPI), which provides a straightforward implementation of the approach depicted in Figure 2. The DPI software leverages and connects different tools, abstracting the different interfaces and implementation details of the underlying tools and applications through simple to use interfaces. The DPI pipelines are available as a CLI tool[6], as a Web Service[7], and include a GUI client application[8].
- OGC Data Exchange Toolkit, which uses Continuous Integration/deployment/testing data workflows. It can process a number of input data formats, apply a combination of simple, atomic transformations on the data, derive linked data representations of it, perform different types of validations on it, entail new metadata and semantic links for it, and store the results in semantic databases.

[6] https://gitlab.pcss.pl/daisd-public/dpi-pipelines/pipelines

[7] https://dpi-enabler-dpi-enabler.apps.dcw1.paas.psnc.pl/api/ (Swagger: https://dpi-enabler-dpi-enabler.apps.dcw1.paas.psnc.pl/api/swagger)

[8] https://dpi-enabler-ui-test.apps.paas-dev.psnc.pl/

- TAPIS (Tables from APIs for Sensors)[9], which allows to annotate a table using an extension of JSON schema or to annotate a CSV table using CSVW (a proposed standard for describing and clarifying the content of CSV tables)

Green Deal Data Space APIs

AD4GD has a strong interest in spatial data and thus it proposed the adoption (and adaptations if needed) of OGC standards. OGC standards define models, formats and exchange mechanisms for localized data defined by the community of practitioners with the long-standing heritage. OGC APIs is the sub-suite of standards modernizing widely adopted OGC Web Services (OWS) like WMS, WCS, WFS, WPS, CSW. They remain compliant with higher level ISO standards and some were adopted as ISO ones. OGC API is a family sharing some principles including OpenAPI descriptions, structure of endpoints, default human and machine-readable encodings. All of them are HTTP request-response services with JSON and HTML encodings on default with various extensions possible. They may be combined in one endpoint with separate relative URLs but also referred from one Landing Page in the fully distributed manner preserving consistency of implementations. Leveraging the OpenAPI capabilities of the self-descriptiveness, OGC APIs standards are accompanied with schema (in JSON schema) and API definition. OGC APIs, particularly relevant for AD4GD, include: OGC API Records (Geospatial data catalogues), SensorThings API (measurements/observations) and OGC OGC API Features (collection of vector features).

AD4GD brings a SensorThings API (STA) generation service deployed on top of the DPI pipelines. It enables access to the Linked Data (in a triplestore) generated by the pipelines via a standard STA interface, which is created automatically from the specified dataset(s) (identified as graph(s)). The dataset(s) should represent observations/measurements that are represented according to the GDIM, i.e. following the SOSA/SSN standard approach. The observations can be grouped in datastreams, and should include (at the observation or collection level) the observed property(ies), feature(s) of interest, sensor(s) making the observation. Also, each observation should include its result (with result and phenomenon time). The service is publicly available online,[10].

[9] https://github.com/joanma747/TAPIS

[10] https://sensor-things-api-sensor-things-api.apps.dcw1.paas.psnc.pl/