Position Paper: Towards Effective Data Interoperability for Data Spaces

Author/Speaker: Thorsten Reitz

Affiliation(s): CEO wetransform GmbH, Co-ordinator FutureForest.ai data space, technical lead for the soilwise data space

Abstract: While there has been some focus and progress on Metadata and Control Plane interoperability, actual data interoperability is essential for most data spaces and still underdefined for many data spaces. However, achieving the right amount of interoperability and re-useability from highly heterogeneous data requires new and effective approaches. In this position paper, we outline our thinking along concrete use cases from the environmental domain.

In the environmental domain, data is traditionally highly fragmented and has had insufficient accessibility. Many data users, controllers and owners have also not fully adopted digital processes. This limits decision and policy making in the light of major challenges such as climate adaptation and nature conservation. Recent work such as the implementation of the INSPIRE directive has improved accessibility and interoperability of non-sensitive data, so that about 10% to 40% of relevant data for environmental use cases is now available, often as harmonised open data. However, as of today, sensitive data such as detailed tree species distributions are usually not findable, accessible, and re-useable outside of the original data controller's domain at all.

As an example, consider the Forestry Data Space that we currently develop in the scope of the FutureForest.ai[1] project. Forest practitioners have to face a massive challenge: Climate is changing very fast, and forest management needs to be adapted to that change. Tried and trusted methods are not reliable anymore, so there is a lot of need for data-driven decision support: When do which forest stands need to be transformed? Which tree species, variants and combinations would thrive in this sport in the future, and would achieve the owner's economic and conservation goals?

New AI-based and data driven tools can really make a difference and help forest practitioners apply specific best practices to their stands, but there is a problem: To be able to apply AI and other tools to this problem at scale, the data used in the process needs to fulfil common requirements. Such requirements cover semantic and structural aspects, as well as quality requirements and the actual encoding, e.g. as an XML- or JSON-based format.

From our perspective, this means that defining these data quality requirements as well as potential standards for the data is a key part of the governance of data spaces. So far, we have mostly seen focus on metadata and on the information required for a data space's control plane, but standardizing the data in the space is essential. Without standardization of the content, a governance model that includes fine-grained rules about who may access what where for which purpose in the data space is impossible to achieve. We need to know that, for example, the

[1] See https://future-forest.eu/

precise location of a protected species is something that is sensitive and may only be processed under certain conditions.

Usually, there are three key challenges to achieving standardization and re-useability of data:

1. The definition or selection of standard taxonomies and vocabularies, i.e. agreeing on common semantics
2. The agreement on minimum data quality requirements and concrete formats and encodings, i.e. agreeing on the actual content and its form
3. Effective onboarding of potentially highly heterogeneous data set from tens of thousands of organisations, i.e. actually populating the data space

For each of these three challenges, there are methods and tools. The actors contributing to the initial development of the governance need to define these and then execute on them, choosing options that integrate well with their communities.

In the environmental domain, there are many well-adopted taxonomies and vocabularies, managed by the respective communities. As an example, the so-called *re3gistry*[2] contains more than 100 vocabularies and taxonomies. Due to this high level of adoption, we have worked to use the re3gistry as the vocabulary provider for the data spaces that we contribute to. However, in the forestry domain, there are no such well-adopted standards outside a few niches. This meant that we had to develop a new formal vocabulary based on so-called ecosystem services models. These ecosystem services cover supportive services, such as biodiversity and contributions to the water cycle, but also provisioning services such as timber production and even cultural services such as recreational value. The agreements on minimal data quality have followed concrete use cases such as the forest transformation and forest vitality analysis.

However, the third aspect, i.e. the effective onboarding of sufficiently harmonised and interoperable datasets into any data infrastructure, has in the past been an area where one of two scenarios was common:

- Specifications and compliance monitoring were too lax, so that data sets are actualy not harmonised and interoperable, but rather still require individual, manual preprocessing for any re-use;
- Specifications and compliance monitoring were too strict, leading to high efforts that did not substantially improve re-usability;

We have thus, across several thousand projects we executed over the last eight years, analysed the factors that contribute most to re-useability and interoperability and compared them to the efforts required. This analysis shows that known data quality, clear documentation and well-defined semantics, early compliance checking as well as the usage of open, well-supported encodings are the factors that have a highly positive cost-benefit ratio.

[2] See https://github.com/ec-jrc/re3gistry and https://inspire.ec.europa.eu/registry

To support such effective measures to onboard data sets and make them interoperable, we are currently developing annotation methods, where new data assets are annotated at different granularity levels (e.g. tables and columns). This process is semiautomated and results in an executable mapping. The declarative system which we are developing builds on top of the open source ETL framework hale studio[3] and the hale connect cloud platform, which enable to automatically transform any data into a common model and encoding based just based on the annotations. For pre-harmonised data sets, i.e. those that already comply with a standard such as INSPIRE, data onboarding into one or more data spaces is actually fully automated.

To summarise, based on our experiences around building data infrastructures at scale and with data from several different environmental domains, we consider it essential for data interoperability to become a key part of the governance and implementation of data spaces. We proposed both methods and tools that can be used and are looking forward to see how other groups are solving these issues.

[3] See https://github.com/halestudio/hale/