CitiObs

Policy Brief 2

2024-11-29

Decision making based on trusted data in the European Data Spaces

A Citizen Generated Air Quality Data example

Executive Summary

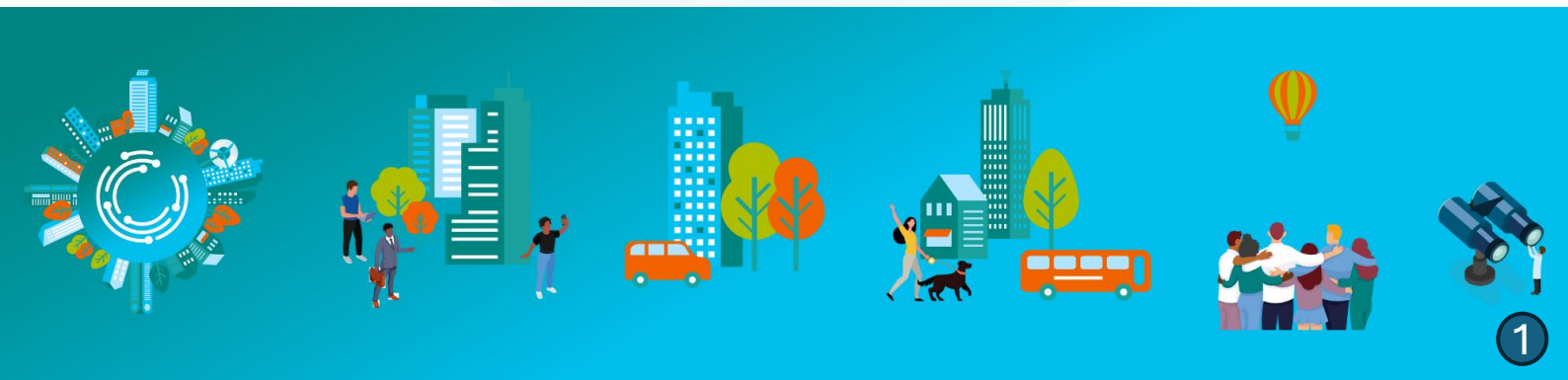
There is a need to transform environmental data into Trusted Decision Ready Information and make it available throughout European Data Spaces. This policy brief presents a solution to establish a single European Data Space as an aggregation of the 15 Data Spaces proposed by the European Commission that includes the Green Deal Data Space.

The proposed solution respects the principles of immutable and discoverable assets, in a truly distributed environment that is technologically neutral, requires minimum modifications in current systems, does not require duplication of assets, and retains the producer data sovereignty.

As a demonstration, CitiObs provides a Citizen Generated Air Quality Data workflow starting with raw data from low-cost sensors for environmental monitoring (e.g. air quality). After validating and integrating it, CitiObs produces Trusted Decision Ready Information at European Scale in a discoverable manner, from an immutable catalog which is subject to certification. The same approach can be generalized to Copernicus Services assets and High Value Datasets (EU 2023/138).

"Summary Points"

1. There is a need for a complementary architecture for Data Spaces based on trusted resources.
2. Data Spaces should exchange Trusted Decision Ready Information.
3. Only validated Citizen Generated Data should be used for decision making.
4. This is a framework on how to achieve trust with minimum changes to existing components.



Introduction

Trustable static validated data assets are needed for contributing to the European Data Spaces. Citizen Science provides data from low-cost sensors, which need validation and integration to be incorporated as in-situ data to improve environmental maps for decision making.

This policy brief provides recommendations on how Citizen Generated Data can be augmented to become a trusted dataset in the Green Deal Data Space.

Acronyms and definitions:
Analysis Ready Data (ARD): Data that has been quality-controlled and allows for immediate analysis with minimal additional user effort.
Decision Ready Information (DRI): Actionable information that can effectively guide decision-making within organizations.
Trusted DRI (TDRI): Certified DRI based on Immutable Catalog and Certificates.
Data Space: Infrastructure that allows participants to exchange data in a trusted environment, promoting collaboration while maintaining control over their data.
Green Deal Data Space: One common European Data Space to support the priority actions of the European Green Deal.

How to make findable and usable Citizen Generated Data

In order to increase trust and uptake of Citizen Generated Data with low-cost sensors, a combination of data validation and aggregation shall take place, resulting in a quality-controlled and validated dataset. This may involve:

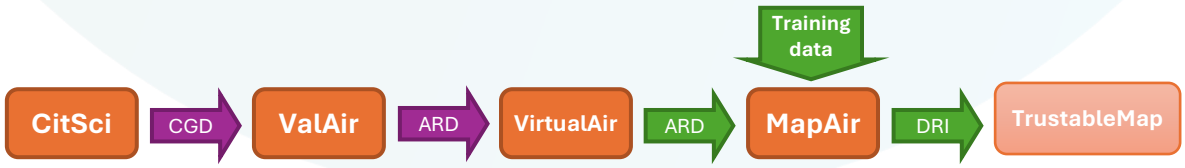
1. automated on-site validation of sensor systems;
2. sensor redundancy (by comparing nearby sensors);
3. data from close reference stations;
4. environmental model outputs;
5. satellite information.

The diversity of low-cost sensor manufacturers and initiatives has increased in recent years. However, before CitiObs, there was no data service that allowed querying quality-controlled Citizen Generated Data from different Citizen Observatories in Europe or worldwide. Without this it was not possible to create an extended analysis.

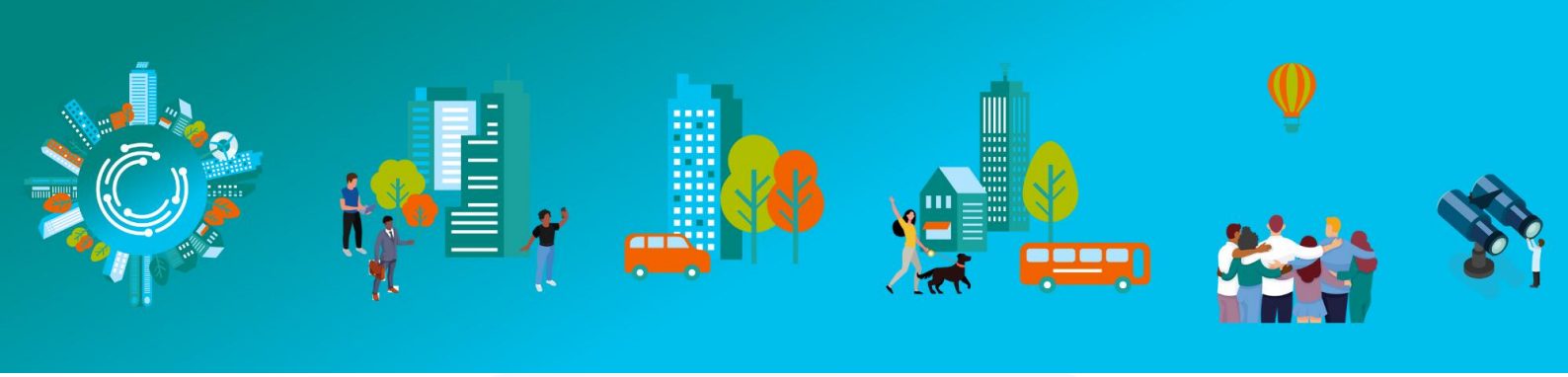
CitiObs addresses these scalability issues by using the Open Geospatial Consortium Sensor Things API PLUS (OGC STAplus) and a set of standard vocabularies to ensure compatibility of the Citizen Generated Data variables along with a service that dynamically aggregates data from different Citizen Observatories in a single queryable virtual dataset. The virtual aggregated dataset (VirtualAir) is respectful of the original data license, propagates data validation levels and adds recognition of the originators.

Decision Ready Information for European Policy requires the production of static map projections generated with trustable snapshots of the Analysis Ready Data available through the VirtualAir.

The next page describes how to make DRI trusted.



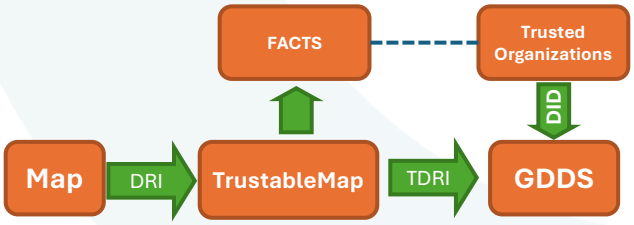
CitiObs



A Green Deal Data Space with Trustable Decision Ready Information

Current approaches to data spaces rely on a technological solution in a *control plane* that connects providers while ensuring digital contracts. This approach forces contributors to set up a technology in their infrastructure to participate in the data space and exchange validated quality-controlled data. However, this does not solve the main problem in data spaces: "Trust in the data without data integration". Alternatively, an approach based on **quality-control, integrity and immutable provenance of datasets** is proposed.

Building upon the existing systems, a trustable dataset can be identified by a unique *hash*, used as a universal identifier. This dataset is registered in an immutable catalog together with its *hash* and the distributed identifier of the data provider. This process provides provenance information and ensures that producers remain in control of their data. Processing in the data space requires a certificate that is issued by the provider.



The certificate may contain license conditions and a list of allowed operations that can be executed using the data. Processing of trustable data in a data space requires a certificate. The execution is only allowed if the certification authorizes the requested processing. The standard specifying this strategy responds to the name FACTS.

Recommendation:
All data and metadata in the Green Deal Data Space should have a *hash* to be registered in an immutable catalog (internally based on a distributed ledgers network). Processing tools should verify the *hash* and get the provenance and request the certification from the provider before processing.

In the CitiObs project we will generate Trustable Decision Ready Information using the MapAir models. However, MapAir capabilities deeply depend on training data therefore, the *training data* should be trusted (i.e. verification of integrity and provenance). The provider of the training data inputs should also be trusted, and the data should be accompanied by a certification that allows the usage or MapAir algorithms. The model generates a result that is catalogued as an immutable asset ready for other participants in the data space to use it.

"Copernicus services and High Value Datasets"

Copying voluminous datasets into an infrastructure is not reasonable. Instead, datasets need to remain in control by the data providers and in the original place of storage.
Recommendation:
•Copernicus services products already have a *hash*. This *hash* should be registered in an immutable catalog to provide discovery of the products and allow certification.
•High Value Datasets produced by the EU administrations[1] should be distributed with a hash registered in an immutable catalog.
Once this is done, they can become part of the Green Deal Data Space.

[1] https://www.copernicus.eu/en/copernicus-services
[2] https://digital-strategy.ec.europa.eu/en/news/commission-defines-high-value-datasets-be-made-available-re-use

“Call to Action”

Consider an alternative approach in the Green Deal Data Space that is based on ensuring trust in environmental data assets based on the principles of immutable and discoverable assets, a truly distributed environment that is technologically neutral, requires minimum modifications in current systems, does not require duplication of assets, retains producer data sovereignty, and can scale up to a single European Data Space.

References

1. D3.2: Final Blueprint of the GDDS Reference Architecture 2024, https://www.greatproject.eu/wp-content/uploads/2024/04/D3.2-Final-Blueprint-of-the-GDDS-Reference-Architecture.pdf
2. OGC Testbed 20: IPT Engineering Report, Paul Churchyard, 2024, https://docs.ogc.org/per/24-033.html

Colophon

Authors:
Dr. Joan Maso (CREAF)
Dr. Andreas Matheus (Secure Dimensions)
Contributor:
Óscar González (IAAC)

How to cite:
Joan Maso, Andreas Matheus, (2024).
Decision making based on trusted data in the European Data Spaces
A Citizen Generated Air Quality Data example. CitiObs policy brief 2.

citiobs.eu @citiobs

CitiObs - Enhancing Citizen Observatories for healthy, sustainable, resilient and inclusive cities

CitiObs

CitiObs is a four-year project funded under Horizon Europe by the European Commission. Over its 48-month duration, CitiObs will collaborate with a total of 85 COs involving members from diverse backgrounds, authorities, scientists, and other stakeholders. Working with established COs, CitiObs strives to create, enhance, and scale up inclusive and diverse observatories, fostering citizen engagement in urban environmental monitoring. The project seeks to expand the deployment of COs as part of inclusive multi-stakeholder governance, generating validated data for policy and research, contributing to GEOSS and GEO initiatives, and co-designing local actions to address urban challenges related to pollution and climate change as part of the European Green Deal transition.

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.