sweet swiss energy research for the energy transition

CROSS

CROSS Data Platform – CROSSDat Concept

Version: CROSSDat-v2022-12

Adriana Marcucci[a]
Christian Schaffner[a]

[a] ETH Zurich

The research published in this report was carried out with the support of the Swiss Federal Office of Energy SFOE as part of the SWEET project CROSS

Cite as:
Marcucci, A., and C. Schaffner. CROSS Data Platform – CROSSDat Concept. Version: CROSSDat V2022-12. SWEET-CROSS.

1 CROSSDat concept

CROSSDat is the data platform of the SWEET-CROSS activity. CROSSDat provides unified and efficient access to SWEET and energy related data, irrespective of where the data is stored and curated.
The main distinguishing feature of CROSSDat is that it is both a database and a metabase that presents data consumers with a platform with aggregated energy data:

- The metabase (database of metadata) automatically understands and harvests metadata from existing external data management systems and databases and presents it. This allows data producers, which already use existing databases, to use CROSSDat without duplicating efforts.
- The database allows researchers from SWEET consortia and beyond to host and directly publish energy-related data (with its accompanying metadata).

Importantly, in CROSSDat, we have a decentralized research data management, which means that the original data providers are responsible of curating the data.

2 Context

Data can play a central role in the transformation of the Swiss energy sector and its digitalization. However, realizing this potential faces some challenges concerning data access and provision. The main challenges regarding data access are high fragmentation and low availability. The very high fragmentation comes from the existence of many data platforms proving energy-related data. A large amount of datasets and databases exist in the energy domain in Switzerland and outside. Researchers, federal offices, cantons, cities, companies have their own preferred data platform, where they have uploaded and will probably continue uploading their research. This lack of an energy-dedicated platform makes access to data inefficient. Additionally, not all data is openly available and it is often exchanged bilaterally. This last aspect of availability is directly linked to the challenges concerning data provision: lack of incentives for data producers; lack of expertise and resources; and distrust in the use of open data due to inexisting or unclear standards and frameworks defining the rights and duties of data users and providers.
CROSSDat aims at tackling some of these challenges, mainly the high fragmentation and the lack of trust. First, to reduce fragmentation, CROSSDat provides an overview of and unified access to energy data, as suggested by the SFOE's report on Open Energy Data (BFE, 2022). CROSSDat allows data providers to continue using their preferred data management system but also offers the possibility of hosting data. Second, we defined the CROSSDat structure following the proposal of the data spaces from the UVEK and EDA (2022), aiming at creating a trustworthy data space.

3 Principles

1. Unified data access: The main objective of CROSSDat is to provide a platform with unified access to SWEET and energy related research data, irrespective of where it is stored and curated.
2. Distributed research data management: Research data management is organized in a decentralized manner where the responsibility to curate research data remains with the experts and the original data providers

3. Findable (and traceable):

    - Datapackages are assigned a unique identifier with a certain standard (e.g. DOI from DataCite) or contain an unique identifier.
    - Data is described with rich metadata
    - Traceable: Different versions should be clearly timestamped and easy to find

4. Accessible:

    - Data can be retrieved by their identifier using a standardized communications protocol
    - No login should be needed to have access to the data. However, data that is not public yet can have restricted access

5. Interoperable:

    - The datapackages use formal, accessible, shared, and broadly applicable language for knowledge representation (frictionless)
    - The platform is able to understand, parse and harvest external sources and present them

6. Reusable:

    - Data includes a clear and accessible data usage license
    - The platform is open source

4 Structure

We defined the structure of CROSSDat following the structures of the data spaces from the UVEK and EDA (2022).

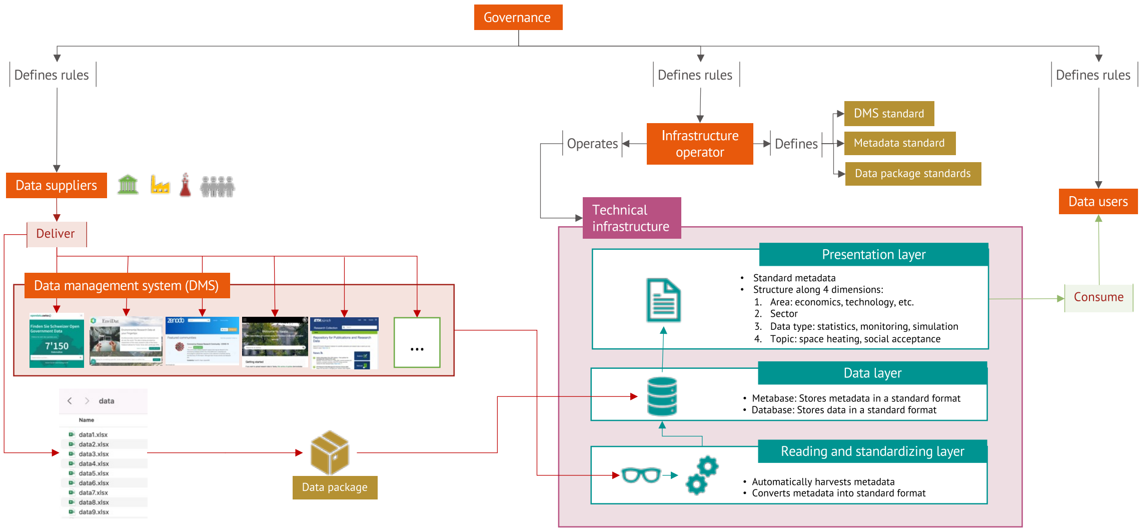

Figure 1: CROSSDat structure

- Data providers deliver data either using their preferred data management system or upload their data to the CROSSDat database.

- The infrastructure operator (1) defines standards for data management systems, metadata and data packages; and (2) operates the technical infrastructure.
- The technical infrastructure includes three layers: (1) Reading and standardizing: Automatically harvest metadata and converts it to the standards; (2) Data layer that stores metadata and data; and (3) Presentation layer that is the interface with the data consumers.
- Finally, on top of all these actors, a governance (still under construction) decides the rights and duties of data providers and data consumers.

5 Core system

CROSSDat uses Frictionless standards. Frictionless standards are based on data packages. Data packages are a format of a container used to describe and package data. Datapackages include a json file describing the datapackage, an example of the structure of a datapackage hosted in the CROSSDat is shown in Listing 1.

Listing 1: Structure datapackage JSON file

```
"file_location":"cross",
"title": "",
"description": "",
"version": "2022-04-05",
"last_updated": "2022-04-05",
"institution": "",
"category": ["",""],
"keywords": [
        "key1","key2"
 ],
"temporal": {
        "start": "",
        "end": "",
        "resolution": ""
    },
"spatial": {
        "location": "",
        "resolution": ""
    },
"documentation":
        {
        "name": "",
        "file": ""
        },
"licenses": [
        {
        "name": "cc-by-4.0",
        "title": "Creative Commons Attribution 4.0",
        "path": "https://creativecommons.org/licenses/by/4.0/"
        }
 ],
 "contributors": [
        {
        "last-name": "",
        "name": "",
        "email": "",
        "affiliation":""
        },
        {
        "last-name": "",
        "name": "",
        "email": "",
        "affiliation":""
        }
],
"resources": [
     {
     "name": "",
     "path": ".csv",
     "mediatype": "text/csv",
     "schema": {
          "fields": [
                    {
                    "description": "",
                    "name": "",
                    "type": "text"
                    },
                    {
                    "description": "",
                    "name": "",
                    "type": "text"
                    }
          ]
          }
     },
     {
     "name": "",
     "path": ".csv",
     "mediatype": "text/csv",
     "schema": {
          "fields": [
                    {
                    "description": "",
                    "name": "",
                    "type": "text"
                    },
                    {
                    "description": "",
                    "name": "",
                    "type": "text"
                    }
          ]
          }
     }
 ],
"sources": [
        {"name": "", "web": ""},
        {"name": "", "web": ""},
        {"name": "", "web": ""}
        ]
}
```

The first step for any data publication in CROSSdat is the registration of a frictionless datapackage describing the data. This datapackage contains either the data and metadata or the information to a platform from which the metadata will be imported. In the second case, the metadata is imported through an automated process (Figure 1). Currently, the platforms from which CROSSDat imports metadata are zenodo, opendata.swiss and envidat.

In the second step the data package is presented to the user. This datapackage includes all the metadata and the links to the files for download.

6 Metadata

The purpose of metadata is to properly document a data entry and to allow CROSSDat to provide certain functionalities such as searchability and accessibility. The quality of the metadata directly affects these functionalities, therefore, the provided metadata should follow the FAIR (Findability, Accessibility, Interoperability and Reusability) principles (Wilkinson et al., 2016):

1. Metadata are described with a formal metadata language, e.g. eCH-0200 DCAT-AP CH

2. Metadata should contain all core attributes

3. Metadata should include a clear usage license

References

BFE, 2022. Open Energy Data Schweiz - Voraussetzung für digitale Innovation im Energiesystem.

UVEK and EDA, 2022. Schaffung von vertrauenswürdigen Datenräumen basierend auf der digitalen Selbstbestimmung.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The fair guiding principles for scientific data management and stewardship. Scientific Data 3, 160018. URL: https://doi.org/10.1038/sdata.2016.18, doi:10.1038/sdata.2016.18.