14. The NextData archive in the context of Open knowledge

E. Trumpy[1]

[1]CNR-IGG, National Research Council, Institute of Geosciences and Earth Resources, Pisa, Italy

Introduction

The main goal of the NextData project is to produce and provide quantitative information on the status and on past, present and future environmental changes in the Italian mountain regions. The information, the data and the achieved results are made available through digital archives, in the spirit of open data.

A dedicated activity of the project was aimed to implement the archive of the databases from the mountain monitoring networks of the project. In particular, the archive provides a detail frame of the on-going weather and climate changes, the atmospheric composition, cryosphere, surface and underground water resources and on ecosystems and biodiversity. Information on mountain ice cores and sediment cores are also embedded in this digital archive. The data collected in this archive allow the monitoring of present as a base for future scenarios (i.e., numerical simulations and modelling results) and past reconstructions (i.e., archives of data).
The built archive makes available the data in the form of spatial datasets, table data or time series together with their associated metadata. Metadata are important to facilitate data organization, data discovery and even data providing. Metadata can be defined as the second level of data, that are useful to describe and classify other data or digital contents. Metadata become crucial when we face large data repositories, as is the case of the NextData archive.
The NextData archive is built ensuring data interoperability with the most relevant global data collection systems, as those implemented by the Group on Earth Observation (GEO) Global Earth Observation System of Systems (GEOSS), Copernicus, Belmont forum.
The concept we use in the NextData archive construction and implementation is in line with the Open Data and FAIR policies now definitely supported and promoted at national and European level.

14.1 NextData archive system

The NextData archive implemented for the needs of NextData project is hosted in a server running in the CNR computing centre in Pisa. In particular, the archive is served by a dedicated Virtual Machine where a Geonetwork application has been specifically set-up and configured to meet the special needs of the project. Table 1 reports the main features and configurations of the physical machine that host the virtual machine and those of the configured Virtual Machine itself.

Hardware - Software		
	Physical machine	**Virtual Machine**
CPU	Intel(R) Xeon(R) CPU E5-2630 v4 10 cores 25MB Socket 2011 v3	1 CPU

RAM	64 GB	16 GB
OS	Ubuntu 18.04 LTS	Ubuntu 18.04.2 LTS
Storage	4 TB	100 Gb
Firewall	Configured to give the access only on port 22 (ssh), 80 (http standard), 8080 (tomcat standard)	Configured to give the access only on port 22 (ssh), 80 (http standard), 8080 (tomcat standard)

Table 1. *Main Hardware and Software features and configurations of the physical machine and Virtual Machine hosting the NextData archive.*

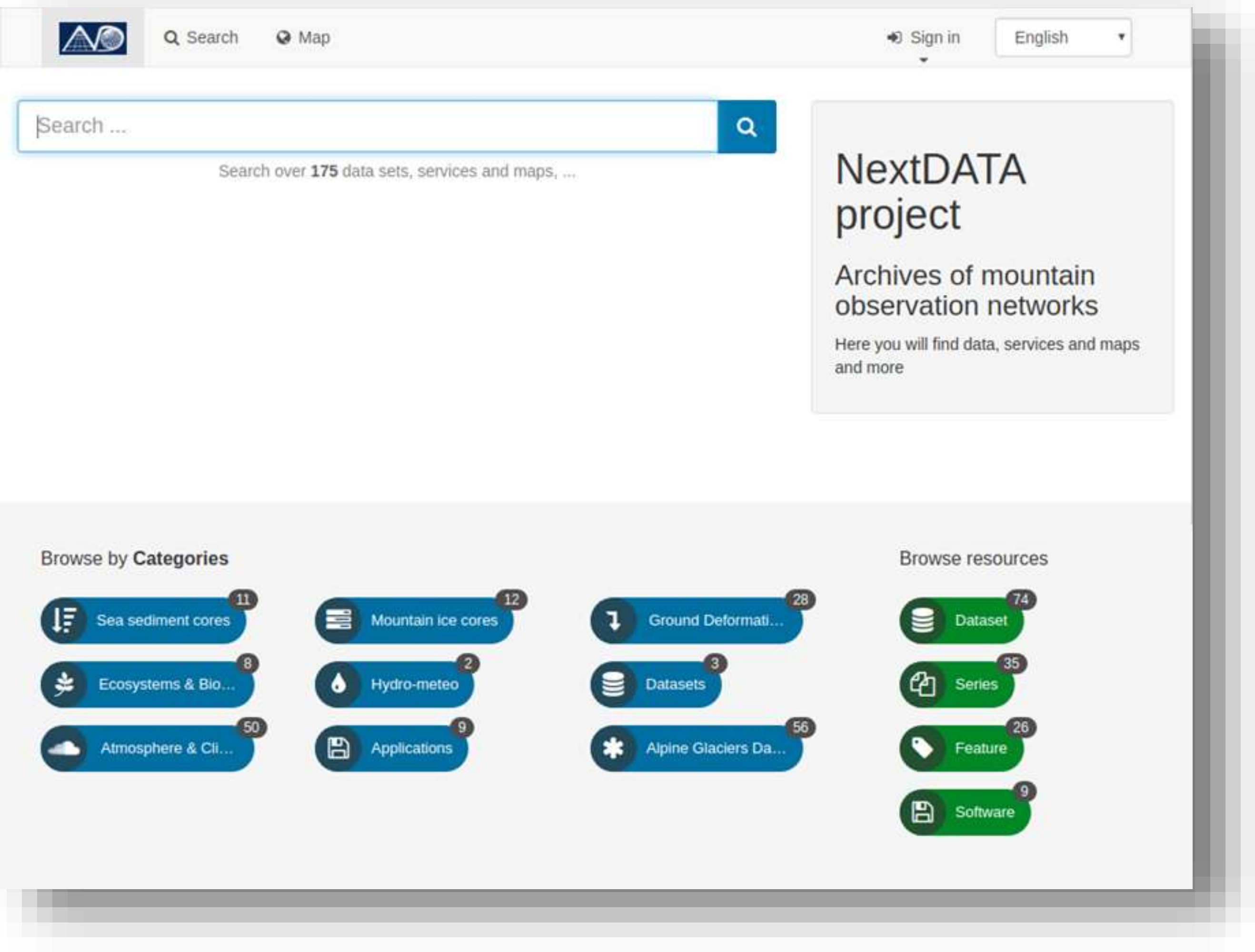
The chosen application to manage the different resources (i.e., datasets, metadata) is Geonetwork at version 3.6. Geonetwork is an open source catalogue specifically developed to manage spatially referenced resources. It catalogues local oriented information and cartographic products through descriptive metadata. Geonetwork enhances the spatial information exchange and share between organizations/departments/projects and their audience by using the capacities and the power of internet (Geonetwork, 2019).

The Geonetwork application implements widely accepted standards to guarantee discovering, viewing and downloading of resources. It exploits Open Geospatial Consortium (OGC, 2019) standards such as dynamic internet map services (e.g., WMS, WFS, WCS), catalogue services (CSW) and makes available different standards to register metadata (e.g., ISO19115, INSPIRE, FGDC, Dublin Core, …).

Currently, 175 metadata are collected and registered by using ISO19115(19139) standard in the NextData archive. The metadata are usually organised hierarchically. It means that, for instance, the dataset produced by a sensor/observation is linked to the metadata describing the sensor/observation itself. The sensor/observation is part of an observatory and in turn the latter is described with another metadata in father-child relationships. In some cases, the metadata father-child dependency was used instead to describe a dataset (father) and its descriptive fields (children). Each resource registered and described with metadata in the NextData archive has the link to download the dataset (i.e., tabular data, grid data, vectorial data, time series data). Where the datasets are not directly available, it is described where the dataset is available and how to download it.

The metadata resources were organized also by categories to facilitate the discovery of the datasets: i) Atmosphere & Climate (50); ii) Alpine Glaciers Database (56); iii) Mountain ice cores (12); iv) Sea sediment cores (11); v) Ground Deformation in Mountain (28); vi) Hydro – Meteo (2); vii) Ecosystems & Biodiversity (8); viii) Applications (9) (in the brackets the number of the current resources registered).

The NextData archive is directly reachable at the URL: http://nextdata.igg.cnr.it or through the official NextData project website (www.nextdataproject.it) looking at 'Data' menu at the 'General Archive' item, Figure 1.



Search over **175** data sets, services and maps, ...

NextDATA project

Archives of mountain observation networks

Here you will find data, services and maps and more

Figure 1. *Screenshot of the NextData digital archive home page.*

14.2 Open Data in NextData

One of the key stones of the Nextdata project is the management of the whole data life. In NextData data are created, described, stored, used and disseminated, in one word Nextdata handles data. There are many reasons why the data management is crucial. Firstly, data can be lost, other data are unique and unrepeatable (e.g. meteorology). Secondly, data have to be organized to foster research activities so that they can improve the research integrity allowing validations and controls. Eventually, managed data could be more visible or favour reuse or even encourage collaboration among scientists.

Data management is an active process by which digital resources remain discoverable, accessible and intelligible over the longer term. The management of data is firstly an internal (e.g., a project, an institution) self-interest, which is the base to get even the benefit for the community externally by providing FAIR and Open data, Figure 2.

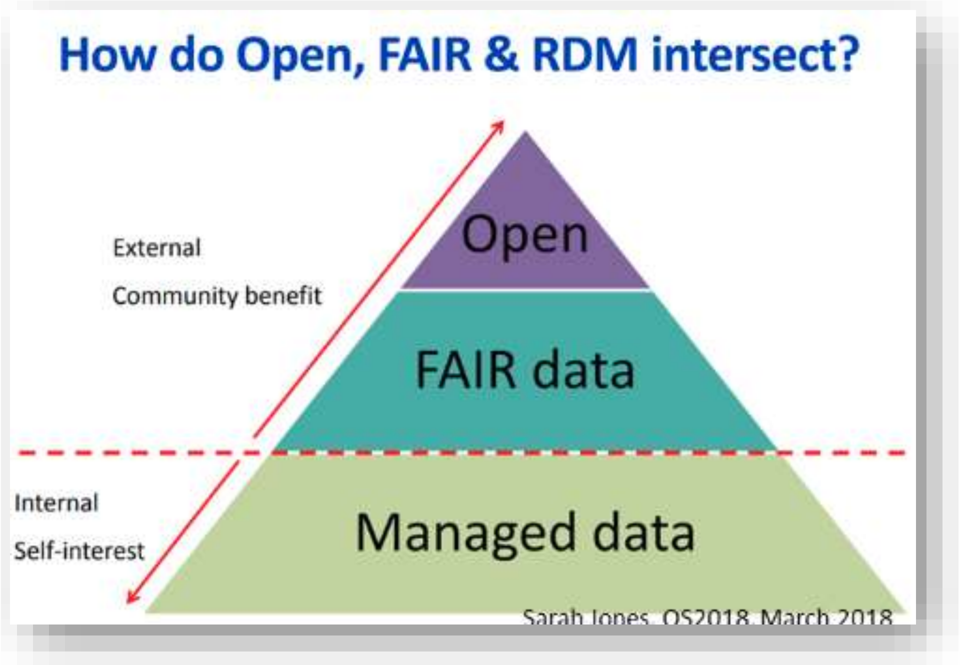


Figure 2. *How do Open, Findable Accessible Interoperable Reusable (FAIR) & Research Data Management (RDM) intersect? (Sarah Jones, OS2018, March 2018).*

In practice ‘Managed data’ means firstly organize data in term of file naming and versioning, then describe them with metadata tackling legal issue as licenses, copyright/left. Eventually, it is crucial the place where store data that it has to allow a long period of conservation.

To favour a right management of data it is highly recommended, even by the European Commission (EC) for the Horizon 2020 funded project, to prepare a Data Management Plan (DMP). The DMP is a document (a deliverable for the H2020 project) describing the data collection, data quality, the storage processes, the way for sharing data and it includes the legal issue and responsibilities on data. The DMP is an alive document, it needs to be updated once important changes occur.

NextData project didn’t produce a DMP because it was not a European project as well as it was designed in the years where Open Data policies and the EU rules were in development. However, the basic idea that was driving the project followed the ‘Internal Self-Interest’ to have ‘Managed data’ (see fig.1). The datasets collected were named clearly and were described with metadata. Moreover, a dedicated repository stores all the data resources guaranteeing a long period of conservation.

The archive of the databases from the mountain observation networks implemented in the project even looks external in order to benefit the community, in particular the scientific one devoted to the climate studies. All data in the NextData digital archive tend to be as much as Open and FAIR as possible.

Open data is data that anyone can access, use and share. The first requirement to yield the data Open is to open it juridically, providing them with an Open Licence so that the user who has the data is free to use, reuse and redistribute even commercially. Aspects like format, structure and machine readability all make data more usable, and should all be carefully considered. Eventually, Open data must be free to use, but this does not mean that it must be free to access since there is often a cost for creating, maintaining and publishing usable data.

Open data can help make research more transparent, facilitate the sharing and reuse (and research validation) as well as reduce the risk to lost data. It can provide the evidence that public money is being well spent and policies are being implemented.

FAIR data could be Open and Accessible is not a synonym of Open. There can be closed FAIR data for security or privacy reasons. Data can be FAIR or Open, both or neither, Figure 3.

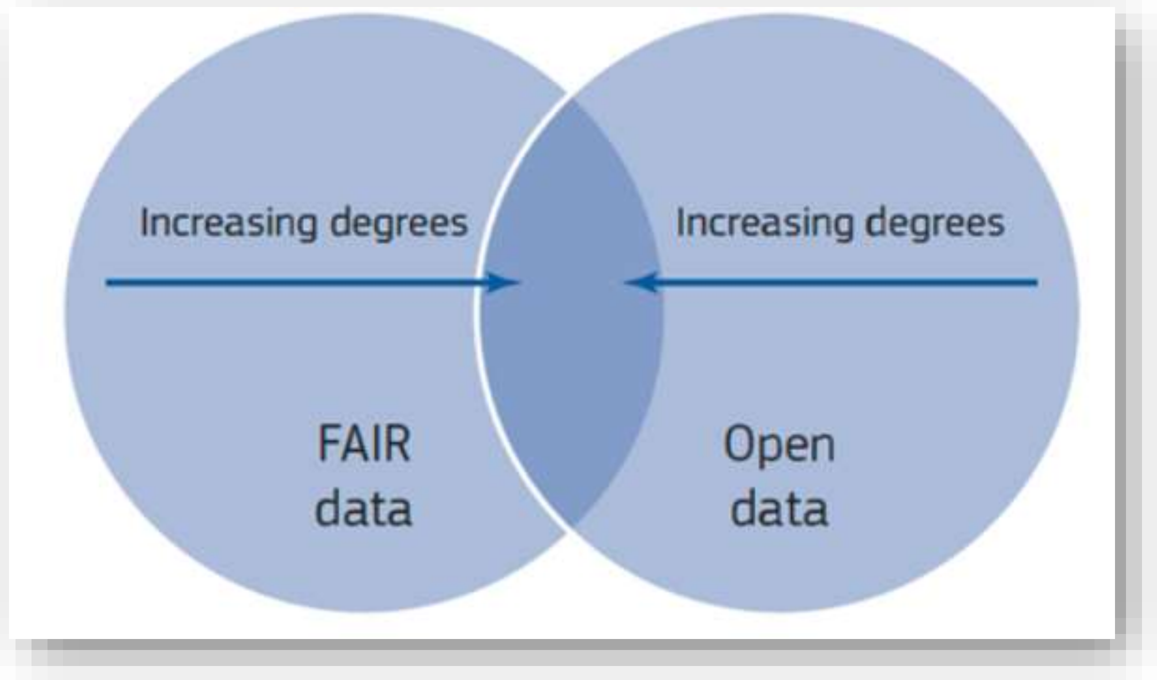


Figure 3. *The relationship between FAIR and OPEN (EC, 2018).*

Open Data have to be firstly Findable, Accessible, Interoperable, Reusable (FAIR). Findable implies that data have to be accompanied with metadata and identifiable/locatable by means of a standard-persistent identification mechanism. Moreover, metadata have to include the identifier of the respective data which in turn must be registered or indexed in a searchable resource.

Accessible is not simply 'Open', the accessibility defines the condition with which data are accessible. To this aim the identifier serves to make data (and metadata) retrievable by using standardized communications protocol (e.g. HTTP, FTP, SMTP, …), which in turn they have to be open, free, universally implementable and allow authentication or authorization procedure where necessary. Although, for some reason (e.g., degrade, cost of maintaining) data could not be no longer available, the metadata must exist to follow the principle of accessibility.

To guarantee the Interoperability of data they should be readable for machine without the need for specialised or ad-hoc algorithms, translators and use controlled vocabularies (also FAIR in turn), ontologies, thesauri and good data model. In addiction is envisaged to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge.

For reusing data, it is recommended to describe specific metadata attributes in order to provide information on the dataset such as the scope, any particularities or limitations, how the dataset has been produced, labs conditions, the provenience and the used community standards (e.g., type of data, data organised in a standardised way, well established file format, documentation, vocabulary, …). To assign a licence on data is another crucial aspect influencing the reuse of data so that MIT or Creative Commons rules of usage are highly recommended.

The datasets in the NextData archive almost satisfied all the requirement to be considered Open and FAIR. The datasets indeed are linked to metadata that are identified by mean of a unique ID. As said, the metadata describe precisely the datasets in agreement with the F of FAIR. The datasets are accessible by common protocols (e.g., http) and the same NextData archive is available as CSW (OGC, 2019) catalogue compliantly with the A of FAIR. The datasets are provided with interoperable and machine readable format (e.g., csv, netCDF, SHP, …). In many cases controlled vocabularies, at least in the metadata keywords, are used as suggested by the I of FAIR. Eventually, many datasets have a licence of use and the scope, provenience, used standards are often described in the metadata as indicated in the R of FAIR.

Conclusions

Most of the time of a researcher is spent to find, homogenize and prepare data, we know even that the 80% of the data are lost in about 20 years (Gibney E. and Van Noorden R., 2013). It is clear that a good management of data is the base for a good scientific research, so data have to be well preserved, documented in a DMP, provided FAIR and possibly Open. To this direction the EC made an important step in the Horizon 2020 and in the next Horizon Europe funding programs. Since the 1st of January 2017 research results, included data, are 'open by default' in agreement with the normative requirements in the EC Model Grant Agreement article 29.2 and in the guidelines (EC, 2016, 2017, 2018). EC promotes also the use of compliant repository to store publications or research data as OpenAIRE or Zenodo.

Open data is only one piece of the wider concept of Open Scienc. Open Science is the movement to make scientific research and data accessible to all for knowledge dissemination and public reuse. All the research workflow can be made Open, so beside the data, the protocols, the software and self-made codes, the methodologies, the publications. To such aims many tools are already freely available, for example repositories for data (e.g. Zenodo, Dryad, Dataverse, …), sharing systems for data, code, notebooks, posters, presentations, protocols and workflows (e.g. Zenodo, GitHub, OpenNotebookScience, protocols.io, FigShare), shared libraries (e.g., zotero), open licences (e.g. CC0 or CC-BY), Open access publications (Green Open Access and Gold Open Access). An Open and free access to text, data, results of the researches guarantee transparency, visibility and reproducibility in the vision of Open Science. Open Science is only the science done right, where 1000 eyes see and solve the problem before two.

At European level is under development the implementation of a federate European Open Science Cloud (EOSC) to guarantee a transparent and cross-disciplinary access to store and work on data, crating a virtual environment where innovators and data producers can meet to boost research and technological development.

In Italy there is the commitment in the PNR (National Research Program) for the realization of a national DataCenter to be linked to the EOSC.

Hopefully, the climate data of the monitoring of present, the data from the past natural archives and the data obtained from the simulation of future scenarios managed by the archive of the NextData project could contribute as federate node firstly to the National DataCenter and in turn to the EOSC.

References

European Commission (EC), Model Grant Agreement – Article 29.2 – Online resources [Last access: 05/06/2019]http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf#page=242.

European Commission (EC), 2016. H2020 Programme Guidelines on FAIR Data Management in Horizon 2020 - Online resources [Last access: 05/06/2019] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

European Commission (EC), 2017. Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 - Online resources [Last access: 05/06/2019] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

European Commission (EC), 2018. Turning FAIR into reality. 2018. Final Report and Action Plan from the European Commission Expert Group on FAIR Data. Directorate General for Research and Innovation Directorate B – Open Innovation and Open Science Unit B2 – Open Science - doi:10.2777/54599.

Geonetworks, 2019. Geonetwork Open Source web site. https://geonetwork-opensource.org [Last access: 05/06/2019].

Gibney E. and Van Noorden R., 2013. Scientists losing data at rapid rate. Nature International weekly journal of science. DOI: 10.1038/nature.2013.14416.

OGC, 2019. Open Geospatial Consortium website. www.opengeospatial.org [Last access: 05/06/2019].

Sarah Jones, Open Data, FAIR data and RDM: the ugly duckling. Open Sciences Conference, Berlin, 13-14 March 2018. DOI: https://doi.org/10.5281/zenodo.1196631.