Building the information system of the French Critical Zone Observatories network: Theia/OZCAR-IS

Isabelle Braud, Véronique Chaffard, Charly Coussot, Sylvie Galle, Patrick Juen, Hugues Alexandre, Philippe Baillion, Annick Battais, Brice Boudevillain, Flora Branger, et al.

Hydrological Sciences Journal

ISSN: 0262-6667 (Print) 2150-3435 (Online) Journal homepage: https://www.tandfonline.com/loi/thsj20

Building the information system of the French Critical Zone Observatories network: Theia/OZCAR-IS

Isabelle Braud, Véronique Chaffard, Charly Coussot, Sylvie Galle, Patrick Juen, Hugues Alexandre, Philippe Baillion, Annick Battais, Brice Boudevillain, Flora Branger, Guillaume Brissebrat, Rémi Cailletaud, Gérard Cochonneau, Rémy Decoupes, Jean-Christophe Desconnets, Arnaud Dubreuil, Juliette Fabre, Santiago Gabillard, Marie-Françoise Gérard, Sylvain Grellet, Agnès Herrmann, Olivier Laarman, Eric Lajeunesse, Geneviève Le Hénaff, Olivier Lobry, Antony Mauclerc, Jean-Baptiste Paroissien, Marie-Claire Pierret, Norbert Silvera & Hervé Squividant

To cite this article: Isabelle Braud, Véronique Chaffard, Charly Coussot, Sylvie Galle, Patrick Juen, Hugues Alexandre, Philippe Baillion, Annick Battais, Brice Boudevillain, Flora Branger, Guillaume Brissebrat, Rémi Cailletaud, Gérard Cochonneau, Rémy Decoupes, Jean-Christophe Desconnets, Arnaud Dubreuil, Juliette Fabre, Santiago Gabillard, Marie-Françoise Gérard, Sylvain Grellet, Agnès Herrmann, Olivier Laarman, Eric Lajeunesse, Geneviève Le Hénaff, Olivier Lobry, Antony Mauclerc, Jean-Baptiste Paroissien, Marie-Claire Pierret, Norbert Silvera & Hervé Squividant (2020): Building the information system of the French Critical Zone Observatories network: Theia/OZCAR-IS, Hydrological Sciences Journal, DOI: 10.1080/02626667.2020.1764568

To link to this article: https://doi.org/10.1080/02626667.2020.1764568

Publisher: Taylor & Francis & IAHS

Journal: *Hydrological Sciences Journal*

DOI: 10.1080/02626667.2020.1764568

Building the information system of the French Critical Zone Observatories network: Theia/OZCAR-IS

Isabelle Braud[a]*, Véronique Chaffard[b], Charly Coussot[b], Sylvie Galle[b], Patrick Juen[b], Hugues Alexandre[c], Philippe Baillion[d,e], Annick Battais[f], Brice Boudevillain[b], Flora Branger[a], Guillaume Brissebrat[e], Rémi Cailletaud[g], Gérard Cochonneau[h], Rémy Decoupes[i], Jean-Christophe Desconnets[j], Arnaud Dubreuil[k], Juliette Fabre[l], Santiago Gabillard[m], Marie-Françoise Gérard[f], Sylvain Grellet[m], Agnès Herrmann[n], Olivier Laarman[b], Eric Lajeunesse[o], Geneviève Le Hénaff[p], Olivier Lobry[k], Antony Mauclerc[m], Jean-Baptiste Paroissien[q], Marie-Claire Pierret[n], Norbert Silvera[r], Hervé Squividant[p]

[a]INRAE, RiverLy, Villeurbanne cedex, France (isabelle.braud@inrae.fr); [b]Univ. Grenoble Alpes, CNRS, IRD, Grenoble-INP, IGE, Grenoble, France; [c]EcoLab, Univ. Toulouse, CNRS, Toulouse, France ; [d]CESBIO, Univ. Toulouse, CNRS, IRD, UPS, CNES, Toulouse, France ; [e]Observatoire Midi-Pyrenées, Univ.Toulouse, CNRS, CNES, IRD, Météo France, UPS, France ; [f]Géosciences Rennes, CNRS, OSUR, Univ. Rennes 1, Rennes, France ; [g]Obervatoire des Sciences de l'Univers de Grenoble (OSUG), CNRS, UGA, IRD, USMB, IFSTTAR, Météo-France, G-INP, INRAE ; [h]GET-UMR CNRS, IRD, UPS–UMR 5563, UR 234, Toulouse, France ; [i]OSU-Réunion - Observatoire des Sciences de l'Univers de La Réunion, La Réunion, France ;[j]ESPACE-DEV, IRD, Maison de la télédétection, Montpellier, France ; [k]LISAH, Univ Montpellier, INRAE, IRD, Montpellier SupAgro, Montpellier, France ;[l]OSU OREME, CNRS, IRD, Univ. Montpellier, Montpellier, France; [m]BRGM, Orléans, France ; [n]Univ. Strasbourg, CNRS, LHyGeS, EOST, Strasbourg, France ; [o]Univ. Paris Diderot, CNRS, Sorbonne Paris Cite, IPGP, Paris, France ; [p]INRAE, Agrocampus Ouest, UMR SAS, Rennes, France ; [q]Univ. Orléans, CNRS, BRGM, ISTO, UMR 7327, Orléans, France; [r]iEES-Paris, SU, USPC, UPEC, CNRS, INRAE, IRD, Bondy, France

*Corresponding author INRAE, RiverLy, 5 Rue de la Doua, CS 20244, 69625 Villeurbanne, cedex, France, isabelle.braud@inrae.fr

Abstract The French Critical Zone research infrastructure, OZCAR-RI, gathers 20 observatories sampling various compartments of the critical zone, each having developed their own data management and distribution systems. A common information system (Theia/OZCAR IS) was built to make their *in situ* observation FAIR (findable, accessible, interoperable, reusable). The IS architecture was designed after consultation of the users, data producers and IT teams involved in data management. A common data model based on various metadata standards was defined to create information fluxes between observatories' ISs and the Theia/OZCAR IS. Controlled vocabularies were defined to develop a data discovery web portal offering a faceted search with various criteria, including variables names and categories that were harmonized in a thesaurus published on the web. This paper describes the IS architecture, the pivot data model and open-source solutions used to implement data discovery, and future steps to implement data downloading and interoperability services.

Keywords data portal; FAIR data; OZCAR-RI; information system; land surface; critical zone; *in situ* observations

1 Introduction

Humanity has entered the era of the Anthropocene (Crutzen 2002), and addressing environmental questions is crucial to ensure that our societies can continue to get the water, food and energy they need in a changing environment. This requires being able to document and simulate the evolution of the Critical Zone, the thin pellicle of the Earth from the top of the atmospheric boundary layer down to the un-weathered bedrock (US National Research Council Committee on Basic Research Opportunities in the Earth Sciences, 2001). Observatories providing long-term datasets documenting the

Critical Zone are required to inform and evaluate earth system models. Such observatories acquire data about the atmosphere, the soil, the cryosphere, the land surface, surface water and groundwater and their geochemical characteristics, some of them going back to the 1960s. However, like most of the research data, the collected data are generally not completely findable, accessible, interoperable and reusable (FAIR, Wilkinson *et al.* 2016). Open data is pushed in Europe by the INfrastructure for SPatial InfoRmation in Europe (INSPIRE) directive for spatial data and the Aarhus agreement[1] for environmental data. To be compliant with these directives, data must be permanently and freely accessible on-line, allowing data discovery, visualization and downloading. Open data is expected to enhance new connections between datasets, data mining, and easier use in models, to foster environmental data science (Gibert *et al.* 2018), meta-analyses or comparison between sites.

In the environmental science community, the challenge of data sharing and interoperability has led to the development of cyberinfrastructures like, in the USA, the Hydrologic Information System for hydrological observatories managed by the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI)[2] consortium (Horsburgh *et al.* 2009, 2011), EarthChem system (Lehnert *et al.* 2010) for geochemical data and CZOData (Zaslavsky *et al.* 2011) for the CZO (Critical Zone Observatory) data management system. All these initiatives had to address semantic and syntactic heterogeneity and proposed shared controlled vocabularies for data variables indexation (e.g. Horsburgh *et al.* 2014) and standard information models for the representation of water-related data. One example is the development of the information

[1] http://ec.europa.eu/environment/aarhus/

[2] https://www.cuahsi.org/

model Observation Data Model ODM 1.0 (and 1.1) and its XML implementation WaterML 1.0 (and 1.1) (Horsburgh *et al.* 2008; OGC 2007) for hydrological time series measured at fixed monitoring points. Its extended version WaterML 2.0 was built from the harmonization of several standards under the umbrella of the Open Geospatial Consortium (OGC, 2012). In order to accommodate different types of data, in particular spatially discrete earth observations, and to be compliant with WaterML 2.0, the ODM1.1 was extended to ODM2 (Horsburgh *et al.* 2016; Hsu *et al.* 2017). In Europe, the ENVRI[3] (ENVironmental Research Infrastructures) community is also developing tools to enhance interoperability and data sharing between research infrastructure gathering communities from various disciplines (e.g. Martin *et al.* 2015; Stocker *et al.* 2016). Moreover, they proposed the ENVRI reference model that describes all the life cycle of data. These common efforts will be continued within the newly launched ENVRI-FAIR project (Petzold *et al.* 2019).

Scientists must be convinced that data sharing, beyond being mandatory, may have a positive impact on their research. Indeed, they sometimes remain reluctant to provide their datasets, putting forward different reasons, such as the lack of technical skills or human resources to check the quality of their data and to provide open data. They also argue that collecting and validating data requires time. Therefore, they want to publish prior to sharing their datasets or they request embargo periods (Blume *et al.* 2017). They often blame the lack of traceability of open data, as they do not know who uses their data and for what purpose. They also blame the possible lack of acknowledgement of their work (Tenopir *et al.* 2011, Desai 2016). Allen and Berghuijs (2018) and Blume *et al*. (2018) discuss incentives for data sharing so that the time spent

[3] https://envri.eu/

to document datasets and publish them can be credited to the researchers and contribute to their career, as proposed in the San Francisco Declaration on Research Assessment (DORA, 2013) or in the Leiden Manifesto recommendation (Hicks *et al*., 2015). Open data also raises practical questions about the definition of a dataset, its granularity, its documentation, the juridical status of data (Becard *et al.*, 2016) and technical issues about interoperability between systems often developed independently, the availability of the required expertise for cyberinfrastructure design and maintenance, and of course associated costs.

The initiative presented in this paper builds on existing national and international initiatives in terms of data sharing. This paper presents the methodology and solutions retained to build a common information system (IS) gathering *in situ* data from all the French Critical Zone Observatories of the OZCAR research infrastructure (RI) (Observatoires de la Zone Critique – Applications et Recherches / Critical Zone Observatories – Applications and Research) (Gaillardet *et al.* 2018). The IS was built within the framework of the French earth system Data Terra RI (Huynh *et al.* 2019), which aims to facilitate access to earth system data. Data Terra is composed of four thematic data poles, where Theia[4] is the pole for land surface data. The objective of building the Theia/OZCAR IS was to make the *in situ* data collected in the land surface community more visible and widely used by the scientific community, using the principles of FAIR data. In order to involve the scientific community, the IS was built using the Agile approach (Beck *et al.* 2001) by a team made of information technology (IT) engineers and data producers. This paper presents how the architecture and functionalities of the Theia/OZCAR IS were designed to meet FAIR principles and

[4] http://www.theia-land.fr/en

users' needs. The paper focuses on the implementation of data discovery for sensor time series, the part of the IS that is already operational. Section 2 describes how the future users' community was consulted and how this information was used to design the Theia/OZCAR IS and the web interface. Section 3 presents the IS architecture and technical solutions to implement data discovery. The advantages and limits of the chosen solution are discussed in Section 4, as well as future steps to implement data downloading and interoperability services.

2 Principles and methodology to build the Theia/OZCAR information system

2.1 OZCAR critical zone observatories research infrastructure and the Theia data pole

To be consistent with the European Strategy Forum for Research Infrastructure (ESFRI), national research infrastructures were labelled at the national level. In France, all the observatories studying the critical zone were gathered within the OZCAR-RI that brings together more than 60 research observation sites, organized in 21 pre-existing observatories (see Table A1 in the Appendix). The ADES[5] French operational portal that provides data about groundwater level and quality is also part of OZCAR-RI. The observatories are operated by diverse research institutions and universities. They were initially created to answer specific environmental questions relevant to society and are promoting long-term observation. One of the criteria for labelling and funding was that data had to be publicly available. Thus, the various communities involved in OZCAR-RI had already made a substantial effort to structure their data and share common formats (e.g. de Dreuzy *et al*., 2006 for groundwater *in situ* and simulation data). They

[5] https://ades.eaufrance.fr/

also developed data and/or metadata portals or repositories where data can be accessed and sometimes downloaded (see details in Table A1), and examples that are more detailed in de Dreuzy *et al.* 2006, Branger *et al.* 2014, Fovet *et al.* 2018, Galle *et al.* 2018, Molénat *et al.* 2018, Pierret *et al.* 2018. However, these efforts have generally been conducted independently. This has led to a very heterogeneous situation, with different levels of development and maturity of the systems and a general lack of visibility of data from the entire OZCAR-RI community.

One ambition of the Theia/OZCAR IS was to improve data accessibility and interoperability by making OZCAR-RI data FAIR. This meant making the data visible, findable and easy to explore; allowing their preservation and their citation; favouring their reuse and sharing; and being interoperable with European infrastructures by adopting international standards (controlled vocabularies and OGC compliant web services). For this purpose, it was decided to build a common data/metadata IS, relying on the experience of the scientific and IT teams involved in the network. In addition, OZCAR-RI belongs to other national and European initiatives, such as Data Terra RI in France and European Long Term Ecosystems Research - Research Infrastructure (eLTER-RI) which has been on the (ESFRI) roadmap since 2018. The Theia/OZCAR IS will have to be interoperable with the IS developed by these infrastructures, which means that technical and architecture choices must be compatible. In France, the Theia/OZCAR IS project team is involved in working groups of the Data Terra RI that aims to harmonize practices between the four French data poles in different thematic areas: atmosphere, ocean, deep surface, land surface, and with the biodiversity pole. The Theia land surfaces data pole is a French national inter-agency organization designed to foster the use of *in situ* and remote sensing observations of land surfaces. Theia offers scientific communities and public policy actors a wide range of data, methods and

services. So far, Theia has focused mainly on access to remote sensing products, but it also aims at providing access to *in situ* data, i.e. data collected in the field either in long-term observatories, either through campaigns or through experiments. Given the complexity and richness of the long-term *in situ* data collected in the OZCAR-RI, it was decided to begin the building of the *in situ* Theia Information System with the data from the OZCAR RI and to focus first on time series. In the future, this IS will be extended to other types of *in situ* data, to short term experiments on land surface, and to *in situ* data used to calibrate satellite products. The methods to analyse current practices and design the Theia/OZCAR IS for *in situ* data are detailed in the next sections.

2.2 Analysis of current practices in data management and dissemination in French observatories

At the beginning of the project, the project team performed a "Tour de France" of the various observatories and data centres providing data within OZCAR-RI. The aim was to present the project, to elicit current practices and identify technical skills for the management and dissemination of data from the observatories. The objective was also to collect the users' needs, but also the reticence towards the new IS. Eight meetings were organized. Seven were face-to-face meetings in the towns where several observatories were present, and the last one was a videoconference with all the observatories not met before. Each meeting gathered between 10 to 20 persons, including the five persons leading the project. During each meeting, the Theia/OZCAR IS team presented the project. The various observatories presented the different types of data acquired in their observatory, the organization of data life cycle and the human resources involved in data management. Then, a time was dedicated to discussions and questions where the participants had the opportunity to ask questions, express their wishes and fears about the future IS. The project team also explained how they planned

to involve IT teams and scientists in the building of the IS and how they would be kept informed of the progress of the project. At this stage, participants also had the opportunity to express the potential barriers for their involvement in the project.

The results of these meetings revealed that data management was currently performed either by research data centres (OSU, Observatoire des Sciences de l'Univers), which gather several observatories, or by IT divisions of some research institutes (e.g. BRGM, INRAE). The map in Fig. 1 and Table A1 (Appendix) highlight an existing degree of regional organization in data management amongst observatories, with nine regional nodes and a large community of IT engineers active in data management.

From the analysis of the outcomes of these meetings and the consultation of the observatory web sites, similarities and heterogeneities were identified in terms of data types and dissemination:

- All portals in OZCAR-RI were providing research data with the exception of the ADES groundwater portal that was providing information about groundwater levels and quality for the whole French territory and was primarily designed for operational use (EU Water Framework Directive reporting).

- Most of the produced data were time series observations acquired by sensors in the field, generally at one location. But some observations were collected on more complex geometries like curves or surfaces (e.g. glacier mass balance is computed using drifting sensors providing data on a trajectory). Other types of data were also identified. This included vector and raster maps (land use, land cover, digital elevation model and soil maps); *ex-situ* data with observations made on samples (e.g. soil cores, soil or water samples) and 2D profiles

describing deep soil structure using different geophysical techniques. Some observatories were producing spatialized products derived from a combination of several data sources, such as: interpolated rainfall fields from raingauges (Vischel *et al.* 2011) and reanalyses combining radar rainfall and raingauges (Boudevillain *et al.* 2016); remote sensing products such as soil moisture fields (Pellarin *et al.* 2009); and model results. Additional types of data included datasets describing agricultural practices, such as application of fertilizers or pesticides, surveys, interviews, photos or videos.

- Different levels of data completeness, quality and processing were provided: metadata only *versus* possible downloading of the data; raw data *versus* corrected data; or, sometimes, more elaborate products including simulation results or gap-filled data.
- Access to the data could be open or granted through provision of a login/password. Some observatories were also imposing an embargo on the data (2–3 years) so that data producers would have time to publish and analyse their data before making them public. Some data were not accessible at all, generally because the data portals were still under construction (Table A1, Appendix), or because data were related to people or sensitive information (e.g. contamination by pesticides). A few observatories were applying a licence on their data, but most of them were providing a data policy that defines how to acknowledge the data producers.
- Data formats were heterogeneous amongst observatories, even for similar variables and sensors.
- The granularity of the datasets was also very variable, depending on the original objectives of the observatory. Indeed, a dataset could comprise a single station, a

network of stations measuring the same variable, or all the data collected within one catchment.

- To give access to data, half of the observatories had developed IS based on relational databases, and web interfaces for searching, documenting and downloading the data. The web interfaces sometimes included map interfaces and visualization tools, while other observatories were providing access to files either directly on the web or through an ftp site. Only three ISs were implementing metadata/data exchange interoperability protocols (OGC standards or their own application programming interface, API).
- The level of information provided in the metadata was also heterogeneous. Some observatories were following ISO 19115 / INSPIRE standards for documenting the metadata of datasets, while others had their own criteria.

2.3 Collection of users' needs for the Theia/OZCAR IS

Regarding expectations towards the Theia/OZCAR IS, the "Tour de France" showed that users were primarily interested in searching data through variable names, spatial location and time slots. The organization of the data into datasets was useful for the producers themselves, as datasets gather coherent sets of co-localized variables, but they are often recognized as hardly understandable by external users. Users were interested in metadata, but once identified, they wanted to go further and get the data itself with a unique extraction format. Data producers were also reluctant to manually fill files with information that they have to provide several times in different formats and were expecting automatized procedures for filling metadata. A previous experience – an effort to develop a metadata catalogue for the hydrological observatories of OZCAR-RI – was also often cited as the "example not to follow". This metadata catalogue (André *et al.* 2015) was conceived to respect the INSPIRE norms and was able to harvest

existing ISs, when the ISs were providing the required web services. For the observatories that were not providing the required services, a manual system was proposed to feed the metadata. However, the usefulness of the data portal remained limited due to the heterogeneous definition of the granularity of the datasets that was not harmonized. Moreover, the names of the datasets were not self-sufficient to understand what they contained. Metadata that were not automatically harvested were quickly obsolete, and metadata documentation was incomplete, implying that access to the data portals was not guaranteed. Users also complained about this catalogue having been built without consulting the future users and not responding to their needs, although the portal was achieving a high technical and interoperability level.

In order to refine the Theia/OZCAR IS functionalities, three types of user were identified: scientists that will use the data portal, data producers and funding institutions. Working groups were organized in April 2018, during the annual OZCAR-RI meeting (40 participants), to elicit the needs for the different groups of users. Each group had to answer the questions that are listed in Table 1. To answer them, participants were asked to write their answer on a post-it, without communicating with others. Then, they were asked to read their answers that were put on a paperboard by one facilitator and grouped when they were addressing similar needs. Another facilitator was in charge of taking note of what participants were saying. At the end, the facilitators of both groups presented syntheses of the main outcomes of the groups in a plenary session. The answers to the various questions are listed in Table 1. In addition, users also expressed the need for proper recognition of the work of collecting and making data available. They asked many questions about the use of DOI (Data Object Identifier) for their datasets, DOI being an interesting way to ensure that their work is well recognized.

2.4 Building the Theia/OZCAR IS according to FAIR principles

All the material gathered during the "Tour de France" and the working groups, as well as exchanges at the national level led to the design of the Theia/OZCAR IS that is presented in Fig. 2. The Theia/OZCAR IS was also designed to fulfil FAIR principles. The reasons for choosing this IS architecture are given below.

The Theia/OZCAR IS had to build on the existing IS of observatories (represented by the green rectangles in Fig. 2), as they already represented a substantial effort to organize, manage and document the data. This was also guaranteeing to keep the information up-to-date and to make the best use of local expertise. In addition, data quality was better guaranteed if data remained as close as possible to their producers (Zaslavsky *et al.* 2011) and the responsibility of the data quality and dissemination had to remain in the observatories. Thus, it was chosen to focus on interoperability with existing systems, so that metadata (and data in a near future) could be accessed transparently, whatever the data location, the local IS and data storage choices (e.g. Ames *et al.* 2012). This principle allowed subsidiarity among observatories. The strength and applicability of a subsidiarity model for developing IS between the central and the local level has largely been demonstrated (Salvemini, 2010).

To interface the existing distributed observatories' systems with the Theia/OZCAR IS, it was proposed to set up an information flux. For each observatory, this flux consists of pushing their metadata and data towards the Theia/OZCAR IS in a format implementing a common data model. This so-called "pivot data model", was built based on standards to have all the necessary information to (i) handle different types of data; (ii) make the data FAIR; and (iii) ensure the functional requirements of the system, such as the search criteria in the web interface and the future implementation of standardized web services. This step is shown with the red rectangle

in Fig. 2. Note that to achieve data exchange, each observatory had to develop a script that extracts the information to be transferred from their existing IS to the central Theia/OZCAR IS in a format respecting the pivot data model (see details in Section 3.2).

The Theia/OZCAR IS was designed to provide services to users or interoperable systems. This includes data discovery, exploration and access functionalities to users through a web interface; standardized metadata/data exchange services actionable by machines (e.g. harvestable catalogue service); services that provide the metadata files compliant with the DataCite Metadata Service Group (2019) that can be transmitted to data producers or pushed towards a DOI registration agency. Users (either humans or machines) consuming these services are represented by the blue rectangles in Fig. 2.

The next section provides more details on the Theia/OZCAR IS architecture. The pivot data model and system architecture were built iteratively using the AGILE approach, with frequent exchanges between the development teams, data producers and the future users.

3 Implementing data discovery in the Theia/OZCAR Information System

The current version of the Theia/OZCAR IS only implements a part of the functionalities described in Section 2.4. They are described in this section and include the definition of a controlled vocabulary for variables names, the definition and implementation of the pivot data model for metadata and data transfer, and the building of the web portal. Currently, the Theia/OZCAR web portal only allows data discovery of time series, and data download is not available yet. The discussion section presents how we plan to complement the Theia/OZCAR IS to implement the remaining functionalities presented in Section 2.4.

3.1 Defining a controlled common vocabulary

Consultation of users showed that their main motivation for using the Theia/OZCAR IS was the ability to find information down to the level of available variables with their associated location and detailed time windows. As a second priority, users cited the ability to know the types of sensors and measurement protocols. The collection of variable names given by the observatories showed different choices of names for the same variable (i.e. rain, rainfall, precipitation, liquid precipitation) in French or English, with more or less precision (daily rainfall, hourly rainfall, rainfall, rainfall at 2 m height). Harmonization of the vocabulary was thus necessary to make the data presented by the Theia/OZCAR IS more discoverable. The use of a hierarchized controlled vocabulary was chosen. This vocabulary was built from the categories and variable names of the Global Change Master Directory (GCMD[6]). Launched by NASA in 1987, the GCMD covers subject areas within the earth and environmental sciences. Today it is one of the largest public keywords inventories in the world. Its use is recommended in the French Data Terra IR. As it does not include chemical data, it was enriched using the SANDRE[7] vocabulary for chemical data (the French norm for data reporting for the EU Water Framework Directive) and with additional categories relevant for the OZCAR-RI community such as "surface fluxes" (Fig. 3). The controlled vocabulary consists of hierarchized concepts of variable categories covering the different compartments of the critical zone, where variable names are found at the last level of the hierarchy. A given variable or a category may be related to several compartments

[6] GCMD keywords : https://earthdata.nasa.gov/about/gcmd/global-change-master-directory-gcmd-keywords

[7] SANDRE : Service d'Administration National des Données et Référentiels sur l'Eau, http://id.eaufrance.fr/gpr/41

(e.g. “surface flux” was related both to “land surface” and “atmosphere”, see Fig. 3). The decision on which variable name was retained in the Theia/OZCAR controlled vocabulary was made by the scientists of the project team, and data producers were then consulted to validate the choices. Without yet including all the variable names of the ADES groundwater operational portal, the final number of variable names that are listed on the Theia/OZCAR web interface is around 300, with about 45% related to chemical data. Some of these data are commonly measured (in more than 80% of the observatories), while a large part (60%) are measured by only one observatory that can, however, measure it at several measurement stations and in different countries.

In order to promote semantic interoperability of the data, a mapping was carried out that links the terms of the Theia/OZCAR thesaurus with exact or similar concepts of international thematic thesauri (see Table 2 for details of the thesauri considered). This mapping was done for each variable name and category name. In addition, Theia/OZCAR variable categories and variable names thesauri were formally described using the SKOS (Simple Knowledge Organization System) standard[8]. They were published[9] using the Skosmos[10] system, an open source web-based SKOS browser and publishing tool (Suominen *et al.*, 2015).

Since the beginning of the project, some data producers have changed their variable names to GCMD, but they can keep their own variable names in their IS. The mapping between the data producers and the Theia/OZCAR variable names is performed in two steps. In the pivot data model, data producers have to specify the

[8] Skos: https://www.w3.org/2004/02/skos/

[9] https://in-situ.theia-land.fr/skosmos/theia_ozcar_thesaurus/en/

[10] Skosmos : http://skosmos.org/

variable category to which their variable names relate (by giving the URI of the variable category in the Theia/OZCAR thesaurus). With the help of the variable category (helpful for example to classify if a water related variable is related to karstic water, ground water or surface water), the scientists in charge of the Theia/OZCAR IS project can associate the data producers' variable names with the variable names of the Theia/OZCAR thesaurus. This is peformed using a homemade web interface that is accessible only to the Theia/OZCAR project team. This web interface also allows creating additional variable names and to update the thesaurus using SPARQL (SPARQL Protocol and RDF Query Language) requests.

3.2 Defining a pivot data model for information exchange between observatories IS and the central Theia/OZCAR IS

According to the "Tour de France" and the consultation of users, it was obvious that one single metadata standard would not be sufficient to cover all the system functionalities that were identified. It was proposed to build a common data model based on different standards of metadata, the so called "pivot data model", as the central element of the Theia/OZCAR IS architecture. We considered the ISO19115/INSPIRE and O&M (Observation and Measurement) standards, as they are necessary to set up standardized metadata/data exchange services, and the Datacite[11] standard, as it is used by DOI registration services. We also considered the schema.org standard, as it allows datasets to be referenced by Google Dataset Search, and the DCAT standard, as it allows facilitating interoperability between data catalogs published on the Web. For building the Theia/OZCAR pivot data model, a mapping between these five standards was performed. Specka *et al.* (2019) performed the same kind of mapping for the INSPIRE

[11]DataCite: https://datacite.org/

and DataCite standards. The mapping between the various standards is provided in a dedicated data model documentation directory of the project's Github repository[12]. The global conceptual schema of the pivot data model is provided in Fig. 4. A complete description can be found in the directory of the data model documentation in the project Github repository[13].

The pivot data model includes three main concepts: data producer, dataset and observation. A dataset is a collection of observations. The information describing datasets (what, where, when, who, how, data use and access condition) was based on the ISO 19115/INSPIRE standard[14] and the DataCite standard. These metadata are required to identify the dataset and its "institutional context". Following Cox (2008) cited in INSPIRE MIG (2016), an *Observation* is an action whose result is an estimate of the value of some property of the feature-of-interest, at a specific point in time, obtained using a specified procedure. The information describing the observation data, such as variable name (*ObservedProperty* concept), feature of interest involved in the observation (*FeatureOfInterest* concept), acquisition and processing methods (*Procedure* concept), observation result (*Result* concept) were based on the O&M OGC standard (OGC, 2013). This kind of information is necessary to understand how the data was acquired and to allow its reuse. The *Result* concept refers to the data itself. It contains some metadata on data values (quality flag description, missing value) and the data filename. For an *Observation* that represents time-series measurements (of a property made repeatedly with the same procedure in a feature of interest), the *Result*

[12] https://github.com/theia-ozcar-is/data-model-documentation/tree/master/standard-mapping

[13] https://github.com/theia-ozcar-is/data-model-documentation/tree/master/pivot-data-model

[14] ISO 19115: https://www.iso.org/standard/53798.html

comprises a single .csv file, which contains only one time series of a given variable at a given location. The format of the .csv file was standardized with a common syntax: the data file must have a short metadata header and data values must be formatted in the form of tuples: date, location, value, quality flags.

When designing the pivot data model, the mandatory fields were those mandatory in each standard to which we added fields needed to meet user-specified needs (like geology, climate, or information about funding institutions and producers). Optional fields of the standards were also considered. Each field of the pivot data model was characterized as mandatory, recommended or optional (see details in the Github repository[15]), in order to maximize the amount of information that could be included in the system, without putting too much pressure on the data producers. The richness of the metadata collected via the "pivot data model" facilitates data discovery and thus the implementation of the FAIR principles.

The use of standardized metadata elements for the pivot data model also ensures the future implementation of standard web-services defined by the OGC such as the CSW[16] (Catalog Service for the Web), the SOS[17] (Sensor Observation Service) for time series, as well as services for obtaining DOIs for datasets.

3.3 Technical solutions for building the Theia/OZCAR IS

It was decided to implement the pivot data model in JSON (JavaScript Object Notation, Bray, 2014), where the geographical features implied in observation, such as

[15] https://github.com/theia-ozcar-is/data-model-documentation/blob/master/pivot-data-model/description_champs_JSON_v1.1%20EN.pdf

[16] CSW (Catalog Service for The Web): http://www.opengeospatial.org/standards/cat

[17] SOS (Sensor Observation Service): http://www.opengeospatial.org/standards/sos

the points, lines or polygons are described in the GeoJSON format (Butler *et al.* 2016). For the data files, a .csv format was adopted. The choice of the JSON implementation was motivated by the following reasons. JSON is less verbose than XML. JSON can be read and understood by humans. On the Theia/OZCAR IS, this format was easier to parse, and there are tools for validating the format of the files (JSON Schema, IETF 2019). The MongoDB database was chosen for storing the metadata as it is a schemaless database that facilitates its evolution if new data types have to be included. Moreover, MongoDB database is a no-SQL document-oriented database that is adapted to store, retrieve and manage JSON semi-structured data.

The technical tools deployed to build the Theia/OZCAR IS architecture are shown in Fig. 5. Figure 5(a) illustrates the information flow from the various observatory ISs to the storage system of the Theia-OZCAR IS. Each data producer has to develop a script that extracts and formats the information from their existing IS. Time series measurements must be provided as .csv files and packaged into .zip archives (one per dataset). Data and metadata are uploaded to Theia/OZCAR IS as fat .zip files using the HTTP PUT request method. Once on Theia/OZCAR IS, each data deposit triggers a process that validates the transmitted information and stores the data. The format of the JSON file and the data files are checked for possible errors (using the JSON schema and regular expressions for .csv files). Validation error details are automatically communicated to the data producers so that they can correct their files. Once validated, the content of the JSON files is imported into a MongoDB database from which it can be processed to feed the web interface or to provide web services. The data files are stored in the file system.

This data flux was organized as a continuous workflow. The frequency of update of the JSON files was left to the data producers. When a new version of a JSON file is

transmitted to the Theia/OZCAR IS, previous information is deleted and replaced by the new one in the Theia/OZCAR IS MongoDB database. Therefore, the copy of the information available in the Theia/OZCAR IS is kept up to date and can be considered as a data cache. For dynamic datasets to which a DOI is assigned, it will be necessary to handle versioning of DOIs as recommended by the Research Data Alliance (Rauber et al., 2015) and already proposed by some data repositories like Zenodo[18].

Figure 5(b) illustrates the Theia/OZCAR IS architecture. This consists of four applications, where each application is divided into three components: a frontend, which is the user interface, a backend, which communicates with the database and contains the logic to send the appropriate data back to the client, and a database:

(i) the import module, which integrates the metadata and data submitted by producers into the Theia/OZCAR IS. The frontend part represents the data deposit area hosted in an Apache server and monitored by an iwatch service which detects the data deposit. The backend part represents the import module developed with a Spring Boot Java framework. Metadata is stored in a MongoDB database using MongoDB Java Driver and data files are stored on a file system.

(ii) the metadata portal, which is the web user interface for discovering the data. The frontend application (web interface) is developed using a Vue.js javascript framework and Leaflet.js javascript library for map view. It is hosted on a Nginx webserver. Queries are made to the backend application (back-end-metadata-portal) through REST API (REpresentational State Transfer Application Programming Interface).

[18] https://help.zenodo.org/#versioning

The backend that is a Spring Boot application communicates with the MongoDB database using Spring Data MongoDB.

(iii) the variable association interface, which is a web application used by Theia/OZCAR administrators to associate producer variable names with the variable names of the Theia/OZCAR thesaurus. Its user interface is developed using a Vue.js framework. Queries are made to the backend application through REST API. The backend-association-variable, which is a Spring boot application, communicates with the MongoDB database using Spring Data MongoDB and via SPARQL requests with the Apache Jena triple store, database used for storing the Theia-OZCAR thesaurus; and

(iv) the thesaurus application, which stores and publishes the Theia/OZCAR controlled vocabulary on the web. Theia/OZCAR controlled vocabulary is formally described in SKOS (Simple Knowledge Organization System). The frontend application (web interface) relies on Skosmos (Suominen *et al.*, 2015), an existing tool developed in PHP, which provides a web interface for browsing the thesaurus, and the Apache Jena triple store database.

The application frontends (when there is a web interface) were developed using the Vue.js JavaScript framework, the SemanticUI CSS framework to apply styles and to allow responsive interface layout and the Leaflet library for the map interface with a marker clustering plugin. The backend components were developed using the Spring Boot Java framework that allows rapid building of micro-services applications. Such applications can be easily deployed in containers that are particularly suitable for a

continuous integration / deployment approach (Jaramillo *et al.*, 2016). In order to automate the deployment of the architecture, we packaged the applications in Docker containers and used the Kubernetes container-orchestration system to allow the system to be easily scalable and fault tolerant.

All the components used in the development are open source so that the system does not depend on possible changes in commercial software distribution and costs. Therefore, the components can also be reused by other projects. Vue.js and SpringBoot were already used by another French data pole for their catalogue and, as the Theia/OZCAR IS is connected to the other French data poles, it was recommended to choose the same tools that allows sharing developments.

3.4 Web portal interface

The design of the web portal interface was done to meet ergonomic criteria such as ease of use, consistency and efficiency. It is a single-page application[19] that offers both a list-based and a map-based search. The users' needs in terms of search criteria, listed in bold in Table 3, were implemented in the web interface that is illustrated in Fig. 6. The search criteria appear as faceted components on the left and allow search by variable names, spatial location, time windows, geology, climate, funding institutions and observatories, as well as full text search. The data that meet search criteria are presented both on an interactive map as clustered markers, and in a list in a folded table at the bottom of the interface (bottom of Fig. 6). The list represents a first level of information with a summarized description including the variable name, the location of measurement, the dataset title to which the variable belongs to, and the data producer.

[19] https://in-situ.theia-land.fr/

The full description of a given variable, the dataset to which it belongs and the producer are accessible using a collapsible component in the right part of the list.

4. Discussion

4.1 About the design of the Information System

The design of the Theia/OZCAR IS was guided by the users wishes. They wanted to search for data by type of variables, location and time windows. Standards like ISO19115/INSPIRE did not allow the provision of all the additional information required by the users (like geology, climate), nor some recommended information for DOI declaration. To combine and reuse data from heterogeneous observatories having various objectives, the "*FeatureOfInterest*" O&M concept was particularly relevant to the OZCAR-RI observatories, as it could contain the notion of sites or thematic objects (e.g. watershed, river, and glacier). Therefore, a more flexible solution was required, that was able to accommodate all the needs. The system also had to be able to evolve over time when new categories of variables or observations will have to be included. In the CUAHSI-Hydrological Information System (HIS) (Ames *et al.*, 2012), data fluxes between distributed ISs and the central IS were organized using a common standardized WaterOneFlow web service where the central system is harvesting the local ISs (pulling system). In the OZCAR-RI, human skills required to maintain such systems were not available in all the observatories. Indeed, some observatories did not have a local IS, nor had they set up web services. Therefore, a solution, making the best use of existing capacities and accepted by data producers, was required. The proposed solution, which consists of the implementation of data fluxes based on a common data model in JSON format, associated with data files in .csv format, was accepted, although it would require effort from their side. Data producers found this effort acceptable, given their existing

workload. Furthermore, the use of the pivot data model allowed each observatory to retain flexibility in the way they organized their data in datasets (let us recall that the information included in the pivot data model is organized in datasets), and to select the datasets they want to share. Data producers have to keep the information about their data up to date in their IS and they can choose the frequency with which they push their data to the central Theia/OZCAR IS. However, to ensure that the information available in the Theia/OZCAR is up to date, it is advisable that they automate the transmission of the JSON + data files once they do an update on their IS. The proposed system requires a minimum of local IT skills in each observatory to be able to maintain the scripts implementing the pivot data model. However, in specific cases, this service can be provided by IT engineers who are part of the Theia/OZCAR central IS. In any case, the success of such an initiative requires the allocation of enough human resources for its creation, maintenance and evolution. Therefore, it relies on the long-term support of the institutions. This is, however, rarely the case, as underlined by Flint *et al.* (2017).

In the design of the Theia/OZCAR IS, a "push" model, in which data producers push their information towards the central Theia/OZCAR IS, was chosen. An alternative could have been to create a register and "pull" model instead. Two options were possible: (i) web services that provide read-only access, like in the CUAHSI HIS that uses WaterOneFlow web services); or (ii) a system in which data producers would give direct access to read their database using solutions like ETL (Extract Transform and Load) or homemade scripts. However, not all data producers were able to implement the first option. The second option would not have been accepted for security reasons. In addition, this kind of solution would have been more difficult to maintain for the Theia/OZCAR IT team, as each extraction script would have to be adapted to each data producer, as each local database has its own structure and not all of them implement the

dataset concept. The “push” solution that was finally adopted required more efforts from data producers at the beginning, as they had to write the extraction script to implement the pivot data model in JSON, but it left them with more flexibility in the way they organized their data. In addition, the collective discussions on the names of categories and variables, on the granularity of datasets and on the metadata gave an active role to the data producers. They became aware of the need to document the datasets. They often had to enrich the information about their data and realized that this was necessary for a larger diffusion and use of their datasets. They also became aware of different options to organize datasets and of the interest of implementing standardized web services. A positive outcome of the collective discussions around the design of the Theia/OZCAR IS was the creation of a community amongst IT and scientists dealing with data management within the OZCAR-RI. Members of this community now know each other and is able to share experiences and skills.

The pivot data model was designed to document time series. However, the inventory of all the data collected in the observatories highlighted other kinds of data such as 2D maps, 2D or 3D profiles (e.g. geophysical measurements), soil profiles, soil cores, experiments and even model simulations. The proposed pivot model will be extended to accept metadata related to these kinds of data using the generic O&M classes of our data model. In addition, the GeoJSON format can handle 2D or 3D geometries. However, we have not yet tested it with this kind of data. Furthermore, MongoDB is schemaless, so evolutions are much easier to handle than in standard relational databases. The choice of implementing the pivot data model in JSON has the advantage that JSON objects are directly imported into the MongoDB database without having to be mapped into java objects in the import module code or the backend code.

This will contribute to facilitate evolutions in the pivot data model. It was already extended to enrich time series description without particular difficulty.

The inventory also revealed that some observatories were collecting "ecological" data that were not time series, such as species names. This could require extensions of the pivot data model to document properties of the station such as local land cover, depth of sensors, etc. We will explore the *FeatureOfInterest* O&M concept to accommodate this kind of data. Some OZCAR-RI observatories also collected social science data, like interviews and questionnaires. However, the extension to such kinds of data has not yet been planned, as it would imply data management complexity that is still a research area (Flint *et al.*, 2017). This issue is currently addressed in other projects like the companion Trajectories project[20] with which complementarity and interoperability is being built.

One advantage of the information flux set up between local ISs and the central Theia/OZCAR IS was that it handles both metadata and data, as the structure of the pivot data model was built to easily accommodate the transfer of data themselves (*Result* class in Fig. 4) jointly with all the associated metadata. Seven observatories have already implemented the pivot data model and transferred it to the Theia/OZCAR IS according to the workflow described in Fig. 5(a). The download of time series from the web interface in a common, standardized, self-documented, open and easy-to-use format for the scientific communities is planned (see Section 4.3) but has not been developed yet. Both the .csv files and .nc (NetCDF with CF convention) formats are targeted. For the moment, links to data producers' web portals are provided. Before providing http download of the data files, we need to: (a) transform the data files

[20] https://trajectories.univ-grenoble-alpes.fr/

exchanged with data producers for supplying fully documented and convenient to use .csv or .nc files; and (b) implement users' authentication in order to manage the data policy of data producers that may result in embargos on recent data, as not all data producers fully fulfil open data principles yet.

The Theia/OZCAR IS described in this paper has a certain level of genericity. In France, the harmonized vocabulary and the pivot data model were defined and extended by the working groups gathering the French data poles in order to accommodate data from the atmosphere, ocean, deep surface and remote sensing communities. The publication of the Theia/OZCAR thesaurus on the web using Skosmos ensures that it can be reused by others. Note, however, that, at present, the variable names are associated with a definition of the term only when the definition was present in the associated semantic concept. Further work will be needed to provide a definition of each variable name, as it is essential to facilitate communication to specialists and non-specialists (Venhuizen *et al*. 2019). In this spirit, a translation into French of the vocabulary should also be provided.

4.2 About the design of the web portal interface

The web portal interface was designed based on users' needs and expectations. In particular, users were really demanding a map interface. The design of the Theia/OZCAR interface is quite similar to the HydroShare[21] interface. HydroShare is the latest version of the CUAHSI data portal, which not only includes publication and discovery of data, but also allows sharing of models, for example hydrological models for a given catchment (Xue *et al.*, 2019), supporting both faceted search and map

[21] https://www.hydroshare.org/

search. Regarding the map interface, HydroShare uses Google Maps and Google Map clustering facilities, while Theia/OZCAR IS uses the open-source JavaScript library Leaflet. Regarding the implementation of the faceted and full text search, HydroShare is based upon the SOLR[22] search engine, which is designed to provide an efficient, advanced, full-text and faceted search capabilities. Theia/OZCAR IS uses the capabilities of the MongoDB database, which provides a less efficient search engine and a less convenient faceting system than SOLR but avoids using two different systems to persist the data for ensuring data consistency, and to index the data for efficient search purposes. "Consistency" in database systems means the database is designed such that every read receives the most recent write or an error. The chosen system favours consistency, whereas systems like SOLR or ElasticSearch prefer "availability" (i.e. such that every request receives a (non-error) response, without the guarantee that it contains the most recent write) to "consistency". So far, the scope of the full text search in MongoDB[23] is sufficient for our needs.

Another important point when designing an information system is the speed of responses to queries. The Theia/OZCAR interface was rendered using client computer capabilities, and the MongoDB database was set in replica-set allowing queries to be distributed between different server instances. Unlike the SOLR search engine, this set-up allows database resilience but cannot distribute operation of a single query between different server instances. For the moment, only one-third of the expected datasets are provided to the Theia/OZCAR IS, so the MongoDB database only stores a small fraction of the metadata. It will be important to verify that query performance remains

[22] https://en.wikipedia.org/wiki/Apache_Solr

[23] https://code.tutsplus.com/fr/tutorials/full-text-search-in-mongodb--cms-24835

acceptable when all of the datasets are included. This will also have to remain true when the datasets will include high frequency datasets that are increasingly common in critical zone studies and have started to be acquired within the OZCAR-RI network (e.g. Floury *et al.* 2017). A solution to face these issues could be to rely on a search engine on top of MongoDB database.

So far, it is not possible to know if the system meets users' expectations in terms of user-friendliness, as the data discovery portal has just been put into production. Available feedbacks that led to modifications and improvements of the web interface came from the scientists leading the project and from scientists and IT teams involved in the building of the Theia/OZCAR IS. This first feedback was positive, and we expect that more feedback from data producers will be available soon.

4.3 Future services provided by the Theia/OZCAR IS

Data discovery is implemented in the current version of the web interface. Reusability of data will be ensured through a data download service that will give transparent access to data for users in common export formats, whatever the origin of the data. Two types of format will be proposed for times series: .csv files and .nc (NetCDF) files, with normalized headers designed in collaboration with other French data poles and data producers. Downloadable data formats have not been defined yet for the other data types. In addition, the web interface provides rich metadata, which allows a contextualization of the data enabling its relevance for the user to be judged. In particular, information about the level of confidence in the data quality and description of data acquisition protocols and treatments will be provided. The pivot data model already incorporates fields related to data quality. This was done using an enumeration list (raw data, quality-controlled data and derived product) and a data quality flag defined by data producers (code and description). These fields have been optional so far

but could become mandatory in the future. For time series, data producers also had the option of providing confidence intervals for their data in the data files (e.g. columns providing min-max uncertainty values).This is an advantage as compared to standard big data harvesting where it is difficult to attribute a quality to the harvested data (Cai *et al.* 2015).

The metadata provided by the data producers also describes information on data license and acknowledgements, as well as the wishes of the data producers in terms of embargo or data diffusion. Therefore, user authentication will be necessary to manage access rights to datasets with embargo, or to manage access granted only to some identified users. Authentication will also be required in order to provide statistics about data downloading to data producers, which is a criterion examined by funding institutions. The implementation of the authentication service will be built in synergy with the other French data poles.

Technical interoperability will be guaranteed thanks to the implementation of standardized web services (OGC). In particular, they will be implemented to allow interoperability with other data infrastructures such as the European eLTER RI with which OZCAR-RI shares some types of data, or Theia spatial data infrastructure. Such services will allow the harvesting of the Theia/OZCAR catalogue via the CSW protocol and the data provisioning of time series via SOS protocol. Interoperability with the Theia and Data Terra ISs will be possible thanks to a pivot data model, common to the four French data poles, and relevant for both remote sensing and *in situ* data, which is already under discussion. Data producers will also benefit from these services to increase the visibility of their data to the stakeholders with whom they work, such as regional councils, local authorities and water agencies. *Ad hoc* extraction scripts from the Theia/OZCAR IS will be developed to provide the required information. Semantic

interoperability has already been guaranteed by the use of a common controlled vocabulary formally described in the SKOS language with semantic links to international, commonly-used, domain-specific controlled vocabularies and accessible on the web.

The building of the Theia/OZCAR IS has allowed a community to be built that shares questions about data management. In order to increase the motivation of data producers to share data and the recognition of their work, training will be provided to help them assign DOIs to their datasets, to provide incentives for the publication of data papers (e.g. Nord *et al.*, 2017; Guyomarc'h *et al.*, 2019), and recommendations for licensing of the datasets (e.g. Creative Commons licenses[24] or ETALAB v2.0[25], the French reference license). A first seminar, conducted in February 2019, was held to discuss the definition of datasets and their granularity. Recommendations on how to fill the associated metadata were also given in order to encourage scientists to publish their data. The *FeatureOfInterest* concept of the O&M standard is particularly relevant to the OZCAR-RI observatories, as it can accommodate the notion of sites (or environmental monitoring facility) or thematic objects (e.g. watershed, river, and glacier see INSPIRE recommendations for hydrological objects) that will be added to the faceted search of the web interface.

Another point not yet addressed by the Theia/OZCAR IS is long-term data preservation and archiving. Providing the data in formats that can be archived with appropriate documentation is a service planned in the IS but it will be added once the

[24] https://creativecommons.org/share-your-work/licensing-types-examples/

[25] https://www.etalab.gouv.fr/wp-content/uploads/2017/04/ETALAB-Licence-Ouverte-v2.0.pdf

system is in production. First contacts were made with CINES[26] which proposes archiving solutions for long-term preservation (Massol and Rouchon, 2010; Diaconnu *et al.* 2014). CINES has to prepare archiving of .csv and .nc formats and we should define jointly the required metadata and documentation needed for future reuse of the data. The corresponding information will be extracted from the Theia/OZCAR MongoDB database in a format specific to their procedure.

5 Conclusions

This paper describes a methodology for designing a data infrastructure from existing distributed information systems with different histories and levels of development, allowing existing data to be rendered FAIR, discoverable, more visible and with a common model for implementing information fluxes between local ISs and the central IS (pivot data model). The "Tour de France" conducted at the beginning of the project was necessary to understand the landscape of data management in OZCAR-RI observatories and allow a positive involvement of data producers (both IT teams and scientists). This led the project team to design the Theia/OZCAR IS as a distributed IS, in which data fluxes are established and pushed by data producers towards the central IS. In order to accommodate all the users' needs collected through working groups, and the future web services that will be implemented, several metadata standards were analysed and merged into a common pivot data model that contains all the required and optional information and allows data transfer. With this solution, data producers have to develop and maintain extraction scripts that implement the pivot data model on their local IS. The Theia/OZCAR IS collects the pushed data and is thus continuously

[26] Centre Informatique National de l'Enseignement Supérieur, https://www.cines.fr/en/

updated. The Theia/OZCAR IS was built using a no-SQL MongoDB database that stores metadata used to respond to users' requests on the web interface. The first part of the project allowed data discovery, and future steps will provide data access and interoperability services. The proposed approach can be adapted to other disciplines and/or data portals with a legacy of distributed data management. The codes and documentation are free and available on Github. The Theia/OZCAR IS will be considered as a success if researchers use it to retrieve the latest versions of their own data. The web portal interface presented in this paper must now be tested by users. As we use an Agile development approach, user feedback will be taken into account to improve and enhance the user-friendliness of the IS.

Codes and documentation availability

The material presented in this paper – the class diagram of the pivot data model, the mapping of metadata fields between various standards and an example of JSON script implementing the pivot data model – are released under the CC-BY-SA licence. The codes are available under the GNU Affero General Public License v3[27] licence. The codes and documentation are available in this project's Github repository: https://github.com/theia-ozcar-is. The version described in the paper was tagged v1.1-beta for each repository, code or document. The list of repositories can be found at https://github.com/theia-ozcar-is?tab=repositories. In each repository, the tagged version can be found by clicking on the linked named 1 release (e.g. https://github.com/theia-ozcar-is/data-model-documentation/releases)

[27] https://www.gnu.org/licenses/why-affero-gpl.html

Contribution of the authors

IB and SG are the scientists in charge of the Theia/OZCAR portal project. VC, CC PJ and RC are the engineers developing the Theia/OZCAR data portal. IB, SG, VC, CC and PJ elaborated the first version of the paper. HA, PB, AB, GB, GC, RD, JCD, AD, JF, SG, MFG, SG, AH, OL, EL, GLH, OL, AM, JBP, NS and HS are engineers in charge of data management and information systems and of the implementation of the pivot data model in OZCAR observatories. BB, FB and MCP are scientists in charge of data access for their respective observatories. All the authors have read and contributed to the final version of the manuscript.

Acknowledgements

The authors thank all the participants of the "Tour de France" meetings, and those that contributed to the working groups during the OZCAR Fréjus meeting in April 2018. The authors are grateful to the Theia technical team (N. Baghdadi, A. Sellé, S. Debard) and more generally the interpole technical group led by F. Genova, including Odatis, Form@terre, Aeris, Theia and PNDB. The persons involved in the Trajectories project from Grenoble Alpes University, and in the Zone Atelier Network are also thanked for fruitful exchanges. The work is conducted within the OZCAR-RI, supported by the French Ministry of Research, French research institutions, universities and the French Agence Nationale de la Recherche (ANR). The work is also part of the Theia land data pole of the Data Terra Research Infrastructure. CNRS/INSU and IRD are also thanked for funding the work of the third author.

References

Allen, S. T. and Berghuijs, W. R., 2018. A need for incentivizing field hydrology, especially in an era of open data: discussion of "The role of experimental work in hydrological sciences – insights from a community survey". *Hydrological Sciences Journal,* 63(8), 1262-1265.

Ames, D. P. *et al.,* 2012. HydroDesktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis. *Environmental Modelling & Software*, 37, 146-156.

André, F. *et al.,* 2015. The RBV metadata catalog, EGU General Assembly 2015, 12-17 April, 2015, Vienna, Austria, Vol. 17, EGU2015-5960.

Beck, K. *et al.,* 2001. Manifesto for Agile Software Development. Agile Alliance. http://agilemanifesto.org/ [Accessed 22 March 2019].

Becard, N. *et al.,* 2016. Ouverture des données de la recherche. Guide d'analyse du cadre juridique en France. DOI : 10.15454/1.481273124091092E12, DOI :

10.15454/1.481273124091092E12, available at http://prodinra.inra.fr/record/382263 [Accessed 15 February 2019]

Blume, T., Van Meerveld, I. and Weiler, M., 2017. The role of experimental work in hydrological sciences – insights from a community survey. *Hydrological Sciences Journal,* 62(3), 334-337.

Blume, T., Van Meerveld, I. and Weiler, M., 2018. Incentives for field hydrology and data sharing: collaboration and compensation: reply to "A need for incentivizing field hydrology, especially in an era of open data"*. *Hydrological Sciences Journal,* 63(8), 1266-1268.

Boudevillain, B.*, et al.*, 2016. A high-resolution rainfall re-analysis based on radar–raingauge merging in the Cévennes-Vivarais region, France. *Journal of Hydrology,* 541, 14-23.

Branger, F., *et al.,* 2014. Database for hydrological observatories: a tool for storage, management and access of data produced by the long-term hydrological observatories of Irstea. *Houille blanche*, (1), 33-38.

Bray, T. 2014. The javascript object notation (json) data interchange format (N° RFC 7159), Internet Engineering Task Force (IETF), available at https://tools.ietf.org/html/rfc7159 [Accessed 23 February 2019].

Butler, H. *et al.,* 2016. The geojson format (N° RFC 7946).available at https://tools.ietf.org/html/rfc7946 [Accessed 23 February 2019].

Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002

Crutzen, P. J. 2002. The "anthropocene". *Journal De Physique Iv,* 12(PR10), 1-5. doi:10.1051/jp4:20020447

DataCite Metadata Working Group, 2019. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.3. DataCite e.V. https://doi.org/10.14454/7xq3-zf69

de Dreuzy, J. R., *et al.,* 2006. General database for ground water site information. *Ground Water*, 44(5), 743-748. doi:10.1111/j.1745-6584.2006.00220.x

Desai, A.R. 2016. Editors' Vox: Your Science Is Your (Openly Shared) Data. Eos, Washington, D.C. https://eos.org/editors-vox/your-science-is-your-openly-shared-data [Accessed 23 February 2019].

Diaconu S. *et al.,* 2014. Scientific Data Preservation. REDON Project White Book, 73p. http://hal.in2p3.fr/in2p3-00959072v1 [Accessed 24 February 2019]

DORA, 2013. San Francisco Declaration of Research Assessment. Available at https://sfdora.org/read/ [Accessed 7 December 2019]

Flint, C. G., Jones, A. S. and Horsburgh, J. S., 2017. Data Management Dimensions of Social Water Science: The iUTAH Experience. *Journal of the American Water Resources Association,* 53(5), 988-996.

Floury, P., *et al.,* 2017. The potamochemical symphony: new progress in the high-frequency acquisition of stream chemical data. *Hydrology and Earth System Sciences*, 21(12), 6153-6165.

Fovet, O., *et al.,* 2018. AgrHyS: An Observatory of Response Times in Agro-Hydro Systems. Vadose Zone Journal, 17(1). doi: 10.2136/vzj2018.04.0066

Gaillardet, J.*, et al.,* 2018. OZCAR: The French Network of Critical Zone Observatories. *Vadose Zone Journal,* 17(1). doi:10.2136/vzj2018.04.0067

Galle, S., *et al.,* 2018. AMMA-CATCH, a Critical Zone Observatory in West Africa Monitoring a Region in Transition. *Vadose Zone Journal*, 17(1). doi : 10.2136/vzj2018.03.0062

Gibert, K.*, et al.,* 2018. Environmental Data Science. *Environmental Modelling & Software,* 106, 4-12. doi:10.1016/j.envsoft.2018.04.005

Guyomarc'h, G., *et al.,* 2019. A meteorological and blowing snow dataset (2000–2016) from a high-altitude alpine site (Col du Lac Blanc, France, 2720 m a.s.l.). *Earth Syst. Sci. Data*, 11(1), 57-69.., 2018, 1-18. doi: 10.5194/essd-11-57-2019

Hicks, D., *et al.*, 2015. Bibliometrics: The Leiden Manifesto for research metrics, Nature, 520(7548), 429–431, doi:10.1038/520429a.

Horsburgh, J. S.*, et al.,* 2008. A relational model for environmental and water resources data. *Water Resources Research,* 44(5), W05406. doi:10.1029/2007WR006392

Horsburgh, J. S.*, et al.,* 2009. An integrated system for publishing environmental observations data. *Environmental Modelling & Software,* 24(8), 879-888. doi:10.1016/j.envsoft.2009.01.002

Horsburgh, J. S.*, et al.,* 2011. Components of an environmental observatory information system. *Computers & Geosciences,* 37(2), 207-218. doi: Horsburgh, J. S.*, et al.* 2011. Components of an environmental observatory information system. *Computers & Geosciences,* 37(2), 207-218. doi:10.1016/j.cageo.2010.07.003

Horsburgh, J. S., *et al.,* 2014. Managing a community shared vocabulary for hydrologic observations. *Environmental Modelling & Software,* 52, 62-73. doi :10.1016/j.envsoft.2013.10.012

Horsburgh, J. S., *et al.,* 2016. Observations Data Model 2: A community information model for spatially discrete Earth observations. *Environmental Modelling & Software,* 79, 55-74. doi :10.1016/j.envsoft.2016.01.010

Hsu, L., *et al.,* 2017. enhancing interoperability and capabilities of earth science data uisng the observations data model 2. *Data Science Journal,* 16(4), 1-16. doi:10.5334/dsj-2017-004

Huynh *et al.,* 2019. L'infrastructure de recherche "Pôle de données et services pour le système Terre », à la pointe des techniques d'imagerie et de cartographie numérique, *Annales des Mines - Responsabilité et environnement*, 94(2), 8-13.

INSPIRE Maintenance and Implementation Group (MIG), 2016. D2.9 Guidelines for the use of Observations & Measurements and Sensor Web Enablement-related standards in INSPIRE, Version 3.0, 154 pp. Available at http://inspire.ec.europa.eu/id/document/tg/d2.9-o&m-swe/3.0 [Accessed 15 April 2020]

IETF (Internet Engineering Task Force), 2019: JSON Schema: A Media Type for Describing JSON Documents draft-handrews-json-schema-02. https://tools.ietf.org/html/draft-handrews-json-schema-02, September 2019 [Accessed 24 October 2019]

Jaramillo, D., Nguyen, D. V. and Smart, R., 2016. Leveraging microservices architecture by using Docker technology. ed. SoutheastCon 2016, 30 March-3 April 2016 2016, 1-5.

Lehnert, K., Walker, D., Chan, C. and Ash, J., 2010. EarthChem: Next generation of data services in geochemistry, *Geochimica Et Cosmochimica Acta*, 74, A578-A578.

Martin, P., *et al*., 2015. Open Information Linking for Environmental Research Infrastructures. 2015 Ieee 11th International Conference on E-Science. 513-520.

Massol, M. and Rouchon. O., 2010. Quality insurance through business process management in a french archive. 7th International Conference on Preservation of

Digital Objects. http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/massol-6.pdf [Accessed 24 February 2019]

Molénat, J., *et al.,* 2018. OMERE: A Long-Term Observatory of Soil and Water Resources, in Interaction with Agricultural and Land Management in Mediterranean Hilly Catchments. *Vadose Zone Journal*, 17(1). doi:10.2136/vzj2018.04.0086

Nord, G.*, et al.,* 2017. A high space–time resolution dataset linking meteorological forcing and hydro-sedimentary response in a mesoscale Mediterranean catchment (Auzon) of the Ardèche region, France. *Earth System Sciences Data,* 9(1), 221-249. doi:10.5194/essd-9-221-2017

OGC (Open Geospatial Consortium), 2007. *OGC® CUAHSI WaterML*. OGC Discussion Paper OGC 07-041r1, I. Zaslavsky, D. Valentine, T. Whiteaker, T Editors, 88 pp. Available at https://www.opengeospatial.org/standards/waterml [Accessed 6 December 2019]

OGC (Open Geospatial Consortium), 2012. *OGC® WaterML 2.0: Part 1- Timeseries, Version 2.0*, OGC 10-126r3, P. Taylor Editor, 149 pp. Available at http://www.opengis.net/doc/IS/waterml/2.0 [Accessed 10 October 2019]

OGC (Open Geospatial Consortium), 2013. *OGC abstract specification. Geographic information — Observations and measurements, S. Version 2.0*, OGC 10-004r3, S. Cox Editor, 54 pp. Available at http://www.opengis.net/doc/is/om/2.0 [Accessed 31 March 2020]

OGC (Open Geospatial Consortium), 2016. *OGC® Earth Observation Metadata profile of Observations & Measurements*, OGC 10-157r4, J. Gasperi , F. Houbie , A. Woolf , S. Smolders Editors, 71 pp. Available at http://docs.opengeospatial.org/is/10-157r4/10-157r4.html [Accessed 17 October 2019]

Pellarin, T.*, et al.*, 2009. Soil moisture mapping over West Africa with a 30-min temporal resolution using AMSR-E observations and a satellite-based rainfall product. *Hydrology and Earth System Sciences,* 13(10), 1887-1896.

Petzold, A.*, et al.*, 2019. ENVRI-FAIR – Interoperable environmental FAIR data and services for society, innovation and research, IEEE International Conference on eScience 2019 (eScience2019), San Diego, 24-27, Oct 2019,

Pierret, M. C., *et al.,* 2018. The Strengbach Catchment: A Multidisciplinary Environmental Sentry for 30 Years. *Vadose Zone Journal*, 17(1). doi : 10.2136/vzj2018.04.0090

Rauber, A., Asmi, A., van Uytvanck, D. and Proell, S., 2015. Data Citation of Evolving data. Recommendations of the Working Group on Data Citation (WGDC), Research Data Alliance (RDA), DOI: 10.15497/RDA00016

Salvemini, M. 2010 : The infrastructure for spatial information in the European community vs regional SDI: The shortest way for reaching economic social development. International Conference SDI – Skopje. http://parcebourg.free.fr/sdiconf2010/Salvemini.pdf

Specka, X*., et al.*, 2019. The BonaRes metadata schema for geospatial soil-agricultural research data – Merging INSPIRE and DataCite metadata schemes. *Computers & Geosciences,* 132, 33-41.

Stocker, M., Ronkko, M. and Kolehmainen, M. 2016. Knowledge-based environmental research infrastructure: moving beyond data. *Earth Science Informatics,* 9(1), 47-65.

Suominen, O. *et al.,* 2015 : Publishing SKOS vocabularies with Skosmos. Manuscript available at http://agilemanifesto.org/ [Accessed 22 February 2019].

Tenopir, C*., et al.,* 2011. Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE,* 6(6), e21101. doi:10.1371/journal.pone.0021101

US National Research Council Committee on Basic Research Opportunities in the Earth Sciences. 2001. Basic Research Opportunities in Earth Science, National Academy Press, Washington, D.C.

Venhuizen, G. J*., et al.*, 2019. Flooded by jargon: how the interpretation of water-related terms differs between hydrology experts and the general audience. *Hydrol. Earth Syst. Sci.,* 23(1), 393-403.

Vischel, T*., et al.*, 2011. Generation of High-Resolution Rain Fields in West Africa: Evaluation of Dynamic Interpolation Methods. *Journal of Hydrometeorology,* 12(6), 1465-1482.

Wilkinson, M. D., *et al.,* 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. doi:10.1038/sdata.2016.18

Xue, Z. K., Couch, A. and Tarboton, D. 2019. Map based discovery of hydrologic data in the HydroShare collaboration environment. *Environmental Modelling & Software*, 111, 24-33.

Zaslavsky, I.*, et al.*, 2011. The Initial Design of Data Sharing Infrastructure for the Critical Zone Observatory. ed. *Proceedings of the Environmental Information Management Conference, iEM'2011*, 2011 Santa Barbara, CA, 28-29 September, 145-150.

Figure captions

Figure 1. Map of the OSU (Observatoires de l'Univers) research data centres and institutional data portals that manage the OZCAR-RI data. The logos refer to the observatories listed in Table A1 (Appendix). Solid (blue) outlines refer to existing and functional data portals; dotted (blue) outlines refer to data portals under construction. The inset (bottom right) is the Reunion Island, located in the western Indian Ocean.

Figure 2. Scheme of the principles of the Theia/OZCAR IS. Different data producers (green rectangles) with different types of databases (relational database, ftp repository) exchange information with the central Theia/OZCAR IS. A continuous information flux is organized through the implementation of the pivot data model (red rectangle). Users (either humans or machines, blue rectangles) can consume the services provided by the Theia/OZCAR IS for different purposes.

Figure 3. Categories of variables retained in the controlled vocabulary of the Theia/OZCAR data portal; these mainly belong to the Global Change Master Directory (GCMD) (purple contours). This was enriched using the SANDRE[28] vocabulary for chemical data (boxes with grey contours) and with additional categories relevant for the OZCAR-RI community (boxes with green contours).

[28] http://id.eaufrance.fr/gpr/41

Figure 4. Simplified visualization of the pivot data model. Information in the blue rectangle corresponds to ISO19115/Inspire information, while that in the red rectangle corresponds to the O&M (observation and measurement) standard.

Figure 5. (a) A typical data import workflow of the Theia/OZCAR IS and (b) architecture of the Theia/OZCAR IS, consisting of four applications: (i) the import module; (ii) the metadata portal; (iii) the variable association application; and (iv) the thesaurus application.

Figure 6. Screen capture of the web interface of the Theia/OZCAR information system. The left panel contains the different filters generated by the faceted classification. The map view displays clustered results of the query. Results are also displayed with a reduced set of information in the list view panel. The full set of information of a given result can be displayed using a collapsible component visible on the right part of the interface.

Table 1. Questions addressed during the working groups in April 2018 and main answers. In bold are highlighted the functionalities that are implemented in the first prototype of the portal web interface.

User type	Questions asked to the working group	Main answers
Scientist	Searching for datasets: how?	**Search by variable, time period, spatial location, compartment of the critical zone (atmosphere, surface water, groundwater, soil, cryosphere, biosphere, land surface), keywords (variable, institution, with or without a tree structure).** **In the case of a search by keywords, users want to be able to refine the search, without having to start from the beginning.** Have a GIS interface where results are displayed for instance when moving the mouse above a location. Have short response time for the search.
	Displaying the results of the search: how?	**Display the result on a GIS interface.** **The filtered list of variables must be displayed.**
	Consult metadata: which ones?	Need to get information about data availability in a given time period (have for instance graphs with the percentage of missing data). **Do not display all the metadata first, but only those that are relevant for the search, with a possibility to get more information if needed.** **If possible, get information on the treatment level and data quality.** **The information must allow the user to know to which compartment of the critical zone the data refer to (e.g. soil, surface water, groundwater, etc.)**
	Visualize selected datasets: how?	Be able to visualize data using a graphical interface, with the possibility to show several variables on the same graphic. Possibly show the graphs on the GIS interface. For maps, have the possibility to see a "quick look" in the GIS interface.
	Get the data: how?	Possibility to get the data in various formats: .csv, .nc, .shp. Possibility to get several variables in the same file (for instance for geochemical data where several parameters are measured based on the same sample). Provision of other variables relevant for another variable if the datasets are properly organized by the data producer (e.g. provide radiation budget and other surface fluxes if the user is looking for evapotranspiration).

		Possibility to get data interpolated at several time steps.
	Additional points raised by the discussion	DTM should be accessible also, as well as vectorial data.
Data producer and PI of observatory	Export datasets to the Theia/OZCAR portal: which metadata?	**Importance of including funding institutions in the way the information is organized and stored.** **Have the possibility to attach documents or publications about datasets.** **Have the possibility to include flags about the quality of the data and the level of processing.** Have the possibility to expose auxiliary data that can be relevant for a dataset (like soil hydraulic properties)
	Get statistics on the datasets: which information?	Information on the people that download the data: country, institution, private/public and planned use of the data Number of downloads and identification of the most downloaded datasets.

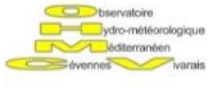
Table 2. Thesauri that were used to map the Theia/OZCAR categories names on existing ontologies. The same thesauri will be used to map the variables names.

Ontology names	Developed by	Field of interest	Web link
GCMD (Global Change Master Directory)	NASA (USA)	Earth Science	https://earthdata.nasa.gov/about/gcmd/global-change-master-directory-gcmd-keywords
AGROVOC	FAO (UN)	Agriculture (FAO)	http://aims.fao.org/fr/agrovoc
GEMET	EEA (European Environment Agency)	Environment	https://www.eionet.europa.eu/gemet/en/themes/
AnaEE	ANAEE-RI (EU)	Continental ecosystems and their biodiversity	https://lovinra.inra.fr/2017/03/13/thesaurus-anaee/
UNESCO	UNESCO (UN)	education, culture, natural sciences, social and human sciences, communication and information	http://vocabularies.unesco.org/browser/thesaurus/fr/
EnvThes	LTER-Europe (EU)	Environment	http://www.enveurope.eu/news/envthes-environmental-thesaurus
LUSTRE Linked Thesaurus fRamework for Environment	eENVplus project (EU)	Environment	http://linkeddata.ge.imati.cnr.it/StartPage.jsp
GACS (Global Agricultural Concept Scheme)	FAO, CAB International and USDA/NAL	Agriculture (FAO)	https://agrisemantics.org/GACS/#browse-gacs-online
NAL (National Agricultural Library)	USDA (USA)	Agriculture (USDA)	https://agclass.nal.usda.gov/
LC Subject Headings Environmental sciences themes	Library of Congress (USA)	Environment	http://id.loc.gov/authorities/subjects/sh92004048.html

Appendix

Table A1. List of the OZCAR-RI observatories, their portal manager (OSU (Observatoire des Sciences de l'Univers), institutional research data centres, or single laboratory), and their characteristics in terms of data management. Observatories are labelled by different French institutions and may gather several distributed observations sites located in France and worldwide (see details in Gaillardet *et al*. 2018)

Logo	Observatory/site and website	Portal manage-ment	Database type	Database portal	Data access	Remarks and link to download the datasets
ORE AgrHyS	AgrHys https://www6.inra.fr/ore_agrhys	Institutional INRAE	Relational	Yes	Public	Metadata and map can be harvested: implements a cataloguing service (CSW webservice) and a map service (WMS web service). Based on GeoOrchestra spatial data infrastructure (Fovet *et al* 2018) https://www6.inra.fr/ore_agrhys_eng/Data
M-TROPICS	M-TROPICS https://mtropics.obs-mip.fr/	Toulouse OSU (OMP)	Relational	Yes	Login/password	https://mtropics.obs-mip.fr/msec-data-access/ https://mtropics.obs-mip.fr/bvet-data-access/
HyBAm	HYBAM http://www.ore-hybam.org/	Institutional IRD	Relational	Yes	Login/password	http://www.ore-hybam.org/index.php/eng/Data
ORE Draix Bléone OTHU	Draix-Bléone https://oredraixbleone.irstea.fr/ ORACLE https://gisoracle.irstea.fr/ OTHU/Yzeron http://www.graie.org/othu/ Real Collobrier	Institutional INRAE	Relational	Yes	Login/password	Common database, BDOH (Base de Données des Observatoires Hydrologique) for the hydrological observatories of Irstea (Branger *et al.,* 2014). Not conceived to easily provide metadata https://bdoh.irstea.fr/

	https://www.irstea.fr/fr/bvre-real-collobrier					
	AMMA-CATCH http://www.amma-catch.org/	Grenoble IGE/OSU (OSUG)	Relational	Yes	Login/password	Fully interoperable. Both metadata and data can be harvested: implements a cataloguing service (CSW webservice) and an data access service (SOS webservice) (Galle *et al.* 2018) http://bd.amma-catch.org
	ERORUN https://osur.univ-reunion.fr/observations/soere/rbv	La Réunion Island OSU (OSU-R)	Under construction	Yes	Login/password	Metadata can be harvested: implements a cataloguing service (CSW webservice) https://geosur.univ-reunion.fr/web/Metadata can be harvested: implements a cataloguing service (CSW webservice)
	OHMCV http://www.ohmcv.fr/	Grenoble OSU (OSUG)	Simple file repository arborescence	Yes	Public	Despite part of the OHMCV observatory data are available in the BDOH and HyMeX databases, the website www.ohmcv.fr is the focal point for all data access
	OHGE http://ohge.unistra.fr/	Strasbourg OSU (EOST) (being discussed)	Simple file repository arborescence	Yes	Ask contact person or login/password	Presently managed by the team of the observatory (Pierret *et al.* 2018) but discussion to be included in the Strasbourg OSU http://bdd-ohge.u-strasbg.fr/index.php/bdd
	ObseRA http://www.ipgp.fr/fr/obsera/observatoire-de-leau-de-lerosion-aux-antilles	OSU (IPGP)	File repository arborescence	Yes	Login/password	WebObsEra portal under construction based on the WebObs IPGP volcanology data portal. http://webobsera.ipgp.fr/

	OMERE https://www.obs-omere.org/	Institutional INRAE, IRD, INGREF, INAT	Relational	Yes	Public	Metadata can be harvested: implements a cataloguing service (CSW webservice) and a map service (WMS web service). Based on GeoOrchestra spatial data infrastructure (Molénat *et al.* 2018) https://www.obs-omere.org/geonetwork/
	Auradé http://www.ecolab.omp.eu/bvea/	Managed by the observatory	Under construction	No	Only metadata	
	SNO Karst https://oreme.org/observation/snokarst/	Montpellier OSU (OREME)	Relational	Yes	All metadata and public data	Metadata and map can be harvested: implements a cataloguing service (CSW webservice) and a map service (WMS web service). Based on GeoNetwork and GeoServer. https://data.oreme.org/observation/snokarst
	SNO H+ http://hplus.ore.fr/en/	Rennes OSU (OSUR)	Relational	Yes	Login/password	Special effort to design a database adapted to a large variety of data including time series, boreholes, soil profiles, experimentation, different types of geophysical data (de Dreuzy *et al.* 2006). http://hplus.ore.fr/base-de-donnees-fr Metadata and map can be harvested through a cataloguing service (CSW webservice) and a map service (WMS web service), based on GeoNetwork and GeoServer. https://www.osuris.fr
	CRYOBS-CLIM https://cryobsclim.osug.fr/	Grenoble OSU (OSUG)	Relational for one part and file repository for the other part	Yes, for part of the data	Login/password	Part of the data are accessible through a portal based on the same information system as AMMA-CATCH. Fully interoperable. Both metadata and data can be harvested: implements a cataloguing service (CSW webservice) and an data access service (SOS

						webservice) http://data.cryobsclim.fr The other part of the data is accessible through files downloading https://glacioclim.osug.fr/spip.php?article75 https://data-snot.cnrs.fr/snot/login.jsf
	Tourbières https://www.sno-tourbieres.cnrs.fr/	Orléans OSU (OSUC)	Relational			
	OSR http://www.cesbio.ups-tlse.fr/fr/osr.html	Managed by the laboratory	Relational and file repository arborescence	Yes	Relational and file repository arborescence	Information system including in-situ data and satellite images and collecting all the data from the laboratory Data portal under construction
	ADES https://ades.eaufrance.fr/	Institutional BRGM	Relational	Yes		Operational database used for the Water Framework Directive reports about groundwater Both metadata, data and map can be harvested: implements a cataloguing service (own API respecting Sandre technical specification), a map service (WFS, WMS web service) and data access service (own API respecting Sandre technical specification) https://ades.eaufrance.fr/
	OPE https://meusehautemarne.andra.fr/landra-en-meusehaute-marne/installations/lobservatoire-perenne-de-lenvironnement	Institutional ANDRA	No	-	Restricted	

Fig 1

Fig 2

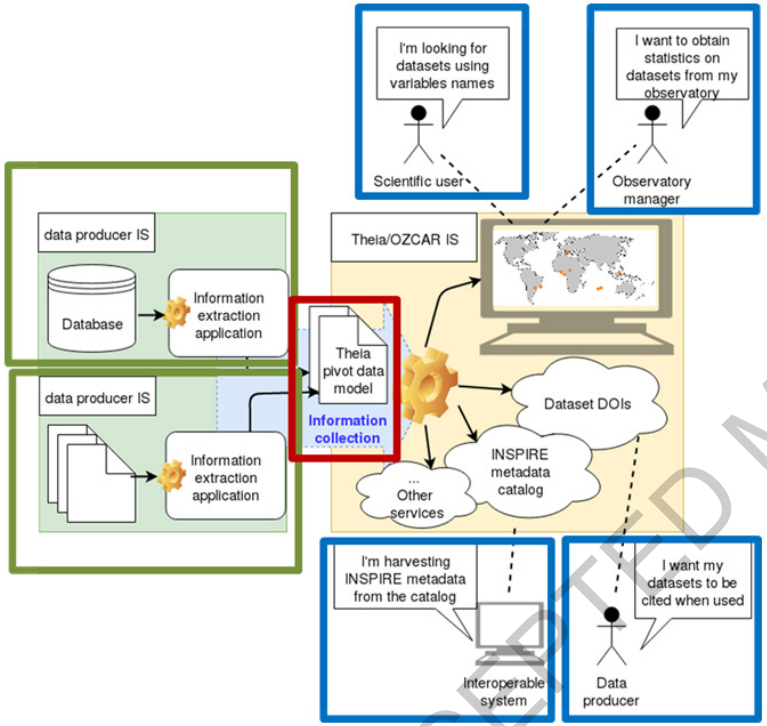


Fig 3

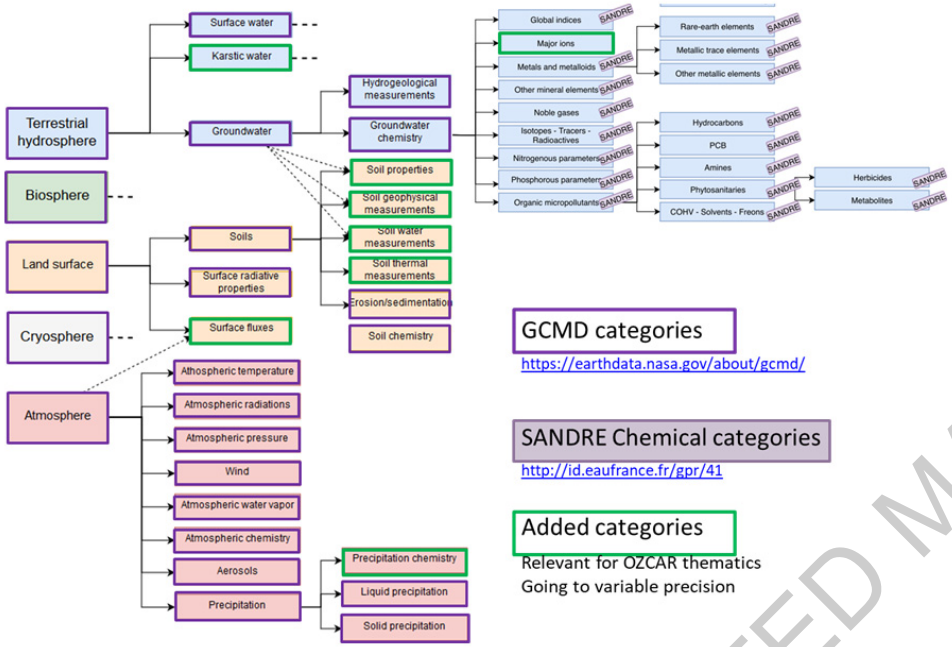


Fig 4

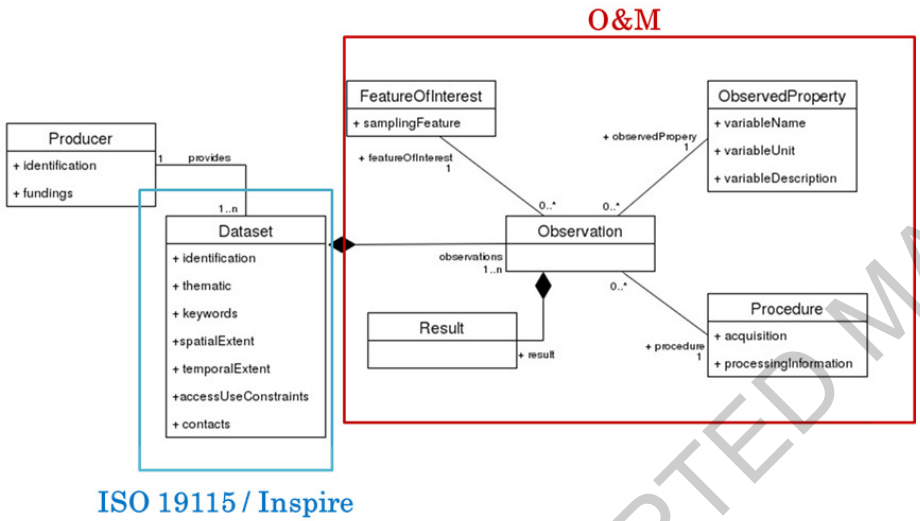


Fig 5

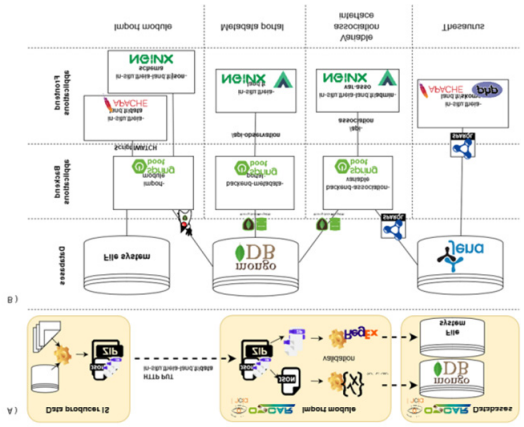


Fig 6