# Enabling FAIR Research in Earth Science through Research Objects

Andres Garcia-Silva[a,*], Jose Manuel Gomez-Perez[a,*], Raul Palma[b], Marcin Krystek[b], Simone Mantovani[c], Federica Foglini[d], Valentina Grande[d], Francesco De Leo[d], Stefano Salvi[e], Elisa Trasati[e], Mirko Albani[f], Cristiano Silvagni[f], Rosemarie Leone[f], Fulvio Marelli[g], Sergio Albani[h], Michele Lazzarini[h], Hazel J. Napier[i], Helen M. Glaves[i], Timothy Aldridge[i], Charles Meertens[j], Fran Boler[j], Henry W Loescher[k], Christine Laney[k], Melissa Genazzio[k], Daniel Crawl[l], Ilkay Altintas[l]

[a] *Expert System, Calle Profesor Waskman 10, 28036 Madrid*
[b] *Poznań Supercomputing and Networking Center PSCN, Jana Pawła II 10, 61-139 Poznań, Poland*
[c] *Meteorological and Environmental Earth Observation MEEO, Viale Volano 195/A Int. 2 I-44123 Ferrara, Italy*
[d] *Istituto di Scienze Marine-Consiglio Nazionale delle Ricerche ISMAR-CNR, Via Gobetti, 101 40129 Bologna Italia*
[e] *Istituto Nazionale di Geofisica e Vulcanologia, Via di Vigna Murata, 605 00143 Roma Italy*
[f] *European Space Agency ESA-ESRIN, Largo Galileo Galilei, 1, 00044 Frascati RM, Italy*
[g] *Terradue Srl, Via Giovanni Amendola,46 00185 Rome Italy*
[h] *European Union Satellite Center, Apdo. de Correos 511, 28850 Torrejón de Ardoz, Madrid – Spain*
[i] *British Geological Survey, Nicker Hill, Keyworth, Nottingham NG12 5GG*
[j] *UNAVCO, Boulder, CO, USA*
[k] *Battelle-National Ecological Observatory, Network, Boulder, CO, USA*
[l] *San Diego Supercomputer Center, UCSD, La Jolla, CA, USA*

**Abstract**

Scientific communities in data-intensive disciplines are progressively adopting FAIR practices that enhance the visibility of scientific breakthroughs and enable reuse. At the core of this transition, research objects contain and describe scientific information and resources in a way compliant with the FAIR principles and sustain the development of key infrastructure and tools in support of research. This paper provides an account of the challenges, experiences and solutions involved in the creation of a FAIR research community, built around the concept of research objects, over several Earth Science disciplines. During this journey, our work has been comprehensive, with outcomes including: an extended research object model adapted to the needs of the different communities of earth scientists; the provisioning of digital object identifiers (DOI) to enable persistent identification and give due credit to authors; the generation of content-based, semantically rich, research object metadata through natural language processing, enhancing visibility and reuse through dedicated recommendation systems and third-party search engines; and various types of checklists that provide a compact representation of research object quality as a key enabler of scientific reuse. All these results have been integrated in ROHub, the research object management platform, which also provides research object management functionality to a wealth of applications and interfaces across the different communities. To monitor and quantify community uptake, we have defined indicators and obtained measures, which are also discussed in the paper.

*Keywords:* FAIR principles, Research Objects, Research Infrastructure, Semantic Technologies, Earth Science

## 1. Introduction

Scientific communities in data-intensive disciplines, together with a diverse group of stakeholders from academia, industry, funding agencies and publishers, are calling for innovative governance models that enhance the visibility of scientific breaktroughs and encourage reuse[1]. Such initiatives seek to overcome the current limitations imposed by conventional scholarly communication as well as

*Corresponding author
*Email addresses:* agarcia@expertsystem.com (Andres Garcia-Silva), jmgomez@expertsystem.com (Jose Manuel Gomez-Perez), jrpalma@man.poznan.pl (Raul Palma), mkrystek@man.poznan.pl (Marcin Krystek), mantovani@meeo.it (Simone Mantovani), federica.foglini@bo.ismar.cnr.it (Federica Foglini), valentina.grande@bo.ismar.cnr.it (Valentina Grande), francesco.deleo@bo.ismar.cnr.it (Francesco De Leo), stefano.salvi@ingv.it (Stefano Salvi), elisa.trasatti@ingv.it (Elisa Trasati), Mirko.Albani@esa.in (Mirko Albani), Cristiano.Silvagni@esa.int (Cristiano Silvagni), rosemarie.leone@esa.int (Rosemarie Leone), fulviomarelli@me.com (Fulvio Marelli), Sergio.Albani@satcen.europa.eu (Sergio Albani), Michele.Lazzarini@satcen.europa.eu (Michele Lazzarini), hjb@bgs.ac.uk (Hazel J. Napier), hmg@bgs.ac.uk (Helen M. Glaves), Timothy.Aldridge@hsl.gsi.gov.uk (Timothy Aldridge), chuckm@unavco.org (Charles Meertens), boler@unavco.org (Fran Boler), hloescher@battelleecology.org (Henry W Loescher), claney@battelleecology.org (Christine Laney), mgenazzio@battelleecology.org (Melissa Genazzio), crawl@sdsc.edu (Daniel Crawl), altintas@sdsc.edu (Ilkay Altintas)

the publication of data[1] and research software[2] in isolated repositories, which hinder scientific progress. Modern science requires to systematically capture the lifecycle of scientific investigations and provide a unified entry point to information about the hypotheses investigated, the data consumed and produced during experimentation or observation, the computations carried out, the conclusions that were derived, the researchers involved in the investigation, and the different licensing models over data or software, to name but a few factors. Some even envision a new grand challenge in science: to create artificial intelligence that can eventually make major scientific discoveries worthy of a Nobel Prize [3]. Still far from realization though, the latter highlights the increasing role of assisted means that support the scientific endeavor.

Research objects are one of the main enablers of such vision, with the potential to accelerate science and estimulate the uptake of good practices in data-intensive science. A research object [4, 5, 6] is a semantically enriched information unit encapsulating all the materials and methods relevant to a scientific investigation, the associated (human and machine-readable) annotations and the context where such resources were produced and came into play. Research objects can be seen as artefacts of both a technical and social nature, with the goal to enhance the sharing, preservation and communication of data-intensive science, facilitating validation, citation and reuse by the community. On the one hand, research objects deal with technical challenges such as preservation, reproducibility, interoperability and platform portability, and contain metadata that make them uniquely identifiable, processable, and exchangeable by machines. On the other hand, research objects attempt to address some of the social aspects crucially involved in the scientific enterprise[7], facilitating that due credit is given to the authors of scientific contributions in their various forms, enabling discussion around the investigation, and ultimately supporting collaboration.

As society and scholars move away from paper towards digital content, research objects have a key role to play in the way scientific results, methods and materials, are communicated, shared, and validated by the scientific community, given the need for mechanisms that support the production and reuse of self-contained, data-centric scientific products. Research objects encourage scientists to share in return of citations to their work, represented as a research object, also shortening publication times. In doing so, research objects support the creation of a virtuous circle of credit and attribution over the key computational resources involved in scientific research. Inspired in sustainable software development practices[8, 9], research objects encourage the release of scientific resources in addition to text publication, in the sense that data, methods and software can be encapsulated as a citable research object in complementary ways to traditional journal or conference publications.

Research objects reinforce the vision contained in the FAIR Data Principles [10], a concise and measureable set of guidelines for those wishing to enhance the reusability of their data holdings. The FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. Data being FAIR is also a way to support the '7-R's' (Reusable, Repurposeable, Repeatable, Reproducible, Replayable, Referenceable, Respectful) that characterize reuse in e-laboratories [4] and initially motivated the creation of the research object concept. The 7-R's fit into the FAIR principles and the desired scientific and research activities in which research objects play the key role. Indeed, as placeholders for everything releated to a scientific investigation, research objects provide a holistic approach towards the reuse of scientific knowledge: not only is data reusable but also put in the context of the investigation as a portable information artefact.

This paper describes the journey of introducing research objects in Earth Science, from the understanding of the needs of these communities in terms of representing, disseminating and reusing scientific knowledge to the required extensions of the research object representation formalism and the associated infrastructure for research object management: ROHub[2]. The work described herein makes special emphasis on the exploitation of natural language processing and semantic annotation technologies to automatically generate research object metadata from their payload, producing richer, self-descriptive, expressive and machine-processable research objects while reducing human annotation effort, thus contributing to FAIR research and its reuse. In this paper we focus on the adoption of research objects by different scientific communities and disciplines in Earth Science, extending previous work in experimental sciences. The paper highlights the role of research objects in making research data FAIR and contextualized in the related scientific investigations and how this has a positive impact in the reuse of scientific knowledge and resources.

The remainder of the paper is structured as follows. Section 2 motivates this work though an analysis of the FAIR-ness level of existing earth observation datasets. Section 3 describes the ecosystem of research object models and tools that we propose in order to enable FAIR research in Earth Science. Section 4 and section 5 describe the research object model, the extensions and customizations introduced to support the specific needs of earth scientists, and the enhancements that support the management of the research life cycle and facilitate sharing through due credit and attribution. Section 6 focuses on the generation of content-based research object metadata, and presents our approach to overcome its scarceness through the application of natural language processing and semantic annotation. Section 7 shows how such metadata is leveraged by dedicated recommender systems and third party

[1] Data Citation Synthesis Group Joint Declaration of Data Citation Principles: https://doi.org/10.25490/a97f-egyk

[2] ROHub is available online at `http://www.rohub.org`

search engines (including bibliographic services), increasing research object visibility within organizations and on the Web, therefore contributing to scientific reuse. Section 8 illustrates how our approach and technologies has been adopted by 3 different scientific communities in Earth Science where scientific data, resources and outcomes are being produced as research objects through dedicated tooling, following the FAIR principles. Section 9 presents our work towards community building. Finally section 10 presents conclusions and future work.

## 2. Motivation

Earth scientists work with heterogeneous datasets generated by data providers such as space agencies, specialized organizations and research projects that produce earth observation data. For example, scientists interested in marine litter need to understand complex scientific inquiries about the distribution and sources of litter, the pathways, the transport mechanisms to the open deep sea, its transformations, the impact on the ecosystem and the sink of marine litter in the marine environment. They work with multiple data types such as: in situ sea floor observations from imaging technology (ROV or Dive transects), fishing trawling, geophysical surveys (e.g. Multi Beam and Side Scan Sonar), visual surveys of floating debris and data for oceanographic modeling.

Were such data published according to the FAIR principles, it would be easier for domain scientists to focus exclusively on the analysis of the data and generate scientific results derived from such observations. However, this is typically not the case. We selected a sample of 35 highly curated, marine research datasets frequently used for marine litter analysis (table 1 shows some of them), collected by public organizations and publicly funded research projects (mainly through EU framework programs and national programs) and assessed their level of FAIR-ness. To this purpose, we followed the methodology proposed by Dunning et al. [11], which systematically evaluates each of the 15 principles corresponding to the 4 letters of FAIR. The methodology considers the information available on the website of the data provider, what is written on help pages, and what is visible in the published data record. The results of our analysis (see figure 1) show that none of the selected datasets can be considered FAIR at the present stage, while most of them do not comply with the FAIR principles. While this analysis only covers a specific area of Earth Science, the conclusions we obtained illustrate the general situation of research data in the observational scientific disciplines.

## 3. Towards FAIR Research in Earth Science

To enable a FAIR research environment we propose an ecosystem of models and tools that interplay in order to help scientists to easily share, find and reuse scientific

Table 1: Shortlist of public marine litter data sets per project

| Project | Format | Size | Period | Area |
|---|---|---|---|---|
| HERMIONE | .shp .csv | 200 KB | 2009-2012 | Artic, Atlantic, Mediterranean |
| PERSEUS | .shp .csv | 200 KB | 2012-2015 | Mediterranean |
| MIDAS | .shp .csv | 200 KB | 2013-2016 | Mediterranean |
| PROMETEO | .mp4 | 1 GB | 2007-2010 | Mediterranean |
| OASIS DEL MAR | .mp4 | 1 GB | 2010-2012 | Mediterranean |
| Ritmare | .shp | 14 MB | 2013 | Venice Lagoon |
| CoCoNet | .shp | 90 GB | 2012-2015 | Adriatic |

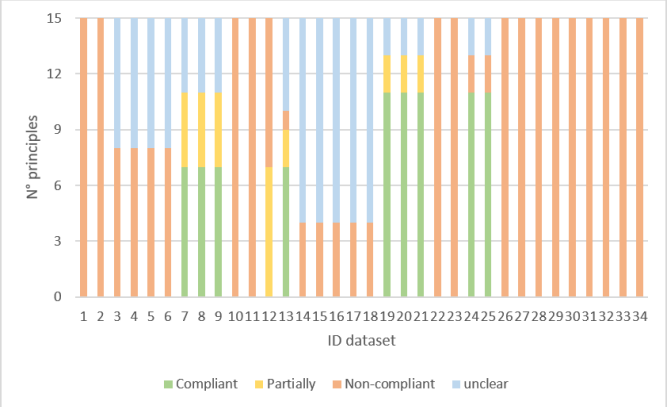


Figure 1: FAIR-ness evaluation of 35 datasets about marine litter

results. The ecosystem depicted in figure 2 contextualizes the contributions presented in this paper in order to enable data-intensive research communities like Earth Sciences to become FAIR.

We argue that the research object model is at the foundations of such transformation, since it defines an agreed vocabulary to share scientific outcomes that makes them interoperable and machine-readable, thanks to the use of a standard data format and a formal semantics. The research object model is generic enough to accommodate any scientific community. Nevertheless, to make it practical for Earth Sciences it must be extended and customized to the specific needs of this area of science. We interviewed and trained earth scientists from research organizations to identify such needs and extended the research object model accordingly.

The research object model enables users to produce metadata about the research object structure, content, and lifecycle. Structure and lifecycle metadata can be generated automatically by a research object management system, like ROHub, assisting the task and relieving scientists from the burden of producing such metadata themselves. However, producing metadata about the content of a research object, e.g. unstructured text like scientific papers and slides among others, is a complex tasks that requires an intelligent management of the information and therefore usually falls on the user side. As a consequence, such key aspect of the metadata related to a scientific investigation is usually neglected and scarce. To generate content metadata we propose a semantic enrichment process that uses natural language processing against the research object payload to

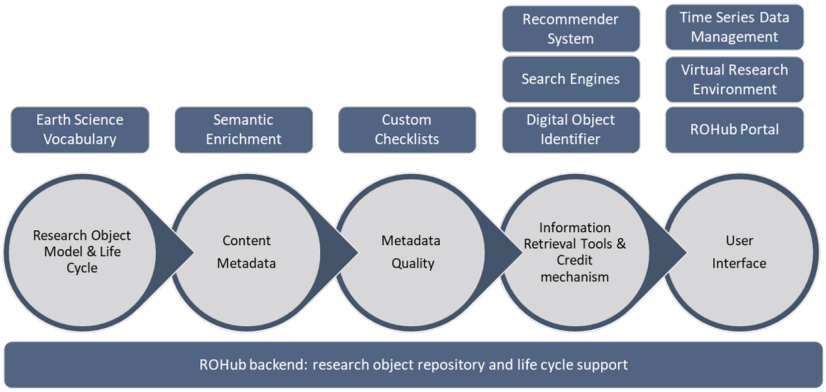


Figure 2: Ecosystem of models and tools based on research objects

generate semantic metadata describing its content.

Rich and expressive metadata is a key factor for sharing and reuse, enabling humans and machines to easily discover relevant research outcomes and related materials as research objects. Therefore, it is necessary to establish mechanisms that monitor their availability and the overall quality of the research object. To this purpose, we provide (and implement) checklists defined in accordance to research object usage scenarios defined in collaboration with earth scientists.

We leverage the semantic metadata generated by our enrichment process, making the research objects in ROHub indexable and searchable by third-party search engines and tools. We have also developed a recommender system that follows a content-based approach to suggest research objects that are similar (in terms of their content) to other research objects selected by a scientist as an input to the system. The interaction with the recommender allows the scientist to engage in a iterative discovery process that leads to the reuse of previous, related work produced by other scientists in the same or another community. Being thus visible and easier to discover, research objects are more likely to be reused by other researchers. In order to ensure due credit is given to the authors of the reused research objects, we have modified the research object lifecycle, extending it with a fork mechanism, inspired from software development practices, that automatically cites the research object being reused. Also, ROHub, now a DataCite[3] member, can assign Digital Object Identifiers (DOI) to research objects upon release of intermediate or final research results.

On top of the model and the metadata, domain-specific applications allow researchers to easily produce and reuse research objects. Here, a core challenge is to integrate research objects with the tools and datasets that earth scientist use in their daily work. At the backbone of such applications, ROHub supports CRUD operations and implements the research object lifecycle. On top of its back-end services and APIs, ROHub offers a generic research object management portal where scientists can create research objects and reuse existing ones from the repository, observing their access policies and licensing schemes. Additionally, earth scientists need specialized user interfaces adapted to earth observation work practices and specific types of relevant data, involving e.g. images, time series, and geolocalized data. In addition to the ROHub portal, in the paper we describe two additional user interfaces working on ROHub's back-end in different communities of earth scientists: a Virtual Research Environment that brings together earth observation datasets and processing tools to do research in Earth Sciences and share outcomes as reusable research objects, and a time series data management application where earth scientist can easily query and visualize on a map real-time data coming from data providers as UNAVCO (GPS data) and NEON (wind and humidity, among others), that can be sliced and stored along with provenance information in a research object.

In table 2 we show how the contributions presented in this paper support data generation according to FAIR principles. The research object models covers practically all the aspects of FAIR, nevertheless, as a model it only enables the generation of FAIR data, but tools that implement the model are required to actually start producing FAIR data. DOI, as permanent identifiers, reinforce findability, and reusability given that they link to metadata about the publication. The semantic enrichment enhances findability by producing rich metadata, while checklists support accessiblity by validating that the metadata is available, and reusability by checking that metadata about license and provenance, to name a few, exists. The visibility of research objects to search engines and the recommender system are another step towards increasing findability. Finally, ROHUB that is complaint to the research object model and integrates the other developments, as well as the virtual research environment and the time-series data management application built upon it, help to generate and reuse FAIR data.

[3] https://www.datacite.org

Table 2: Models and tools in support of knowledge sharing and reuse following FAIR principles

| | Models & tools \ Principles | Research object model + Earth Science Extensions | Digital Object Identifiers (DOI) | Semantic Enrichment & Quality Assessment | Search Engines & Recommenders | User interfaces: ROHub portal | User interfaces: EVER-EST VRE | User interfaces: Time Series Data Management |
|---|---|---|---|---|---|---|---|---|
| F | Rich Metadata<br>(meta)data searchable<br>Persistent Identifier | | | | | | | |
| A | (meta)data retrievable<br>Open & universal protocol<br>Authentication & Authorization | | | | | | | |
| I | Formal Knowledge Rep<br>FAIR Vocabularies<br>link to other metadata | | | | | | | |
| R | usage license<br>provenance<br>Standard community meta(data) | | | | | | | |

## 4. Research Objects

Research objects represent scientific knowledge in a form, rich with annotations, that makes it recognizable, processable, and exchangeable by both humans and machines. A research object is a semantically rich aggregation of resources that bundles together essential scientific information about to a scientific investigation [4]. This information is not limited merely to the data used and the methods employed to produce and analyze such data, but it may also include links to the members of the investigation as well as other important metadata that describe the characteristics, inter-dependencies, context and dynamics of the aggregated resources [4] [5]. As such, a research object can encapsulate scientific knowledge and provide a mechanism for sharing and discovering reusable assets of the investigation within and across relevant communities, and in a way that supports the reliability and reproducibility of the results of such investigation. Nowadays, ROHub [12] is the reference platform for research object management, with myExperiment as its nearest precursor [13].

While there are no pre-defined constraints related to the type of resources that a research object can contain, in the context of scientific research the following usually apply:

- Data used and produced by the experiment or observation.
- Scientific methods applied.
- Software and workflows implementing the methods.
- Provenance and execution settings.
- People involved in the investigation.
- Annotations about these resources, to interpret the scientific outcomes captured by a research object.

The research object model relies on the W3C Resource Description Framework RDF [14], a data model specifically designed for data interchange in the web, and the Web Ontology Language OWL [15], a rich knowledge representation model. In practice, this means that research objects can be easily processed not only by humans but also by machines, since both data and its semantics are described following standard means. The research object model comprises a set of vocabularies that allow describing a research object formally. Such vocabularies are defined in the following ontologies:

- **The Research Object Core Ontology**[4] (ro), describing the aggregation of resources in the research object, as well as the annotations made on those resources.
- **The Workflow Description Ontology**[5] (wfdesc), meant as an upper ontology for more specific workflow definitions, and as a way to express abstract workflows.
- **The Workflow Execution Provenance Ontology**[6] (wfprov), for the representation of provenance information generated by the execution of a scientific workflow.
- **The Research Object Evolution Ontology**[7] (roevo), which describes research object lifecycle information.

Aggregation is supported through the use of the OAI-ORE vocabulary while annotation is supported by the Web Annotation Ontology[8]. In addition, the research object model makes use of existing vocabularies, in particular, Friend of a Friend (FOAF), Dublin Core Terms (DCTerms), and the Citation Typing Ontology (CITO), to provide research object authors with the means to express aspects such as the contributors to a research object, its citations, and the dependencies the research object and its content may have.

Figure 3 shows a graphical representation of an existing research object [9] that uses the core vocabulary. This research object shows a partial and simplified view of the structure of an existing exemplary research object, which uses several modules of the research object ontology suite. It contains a habitat suitability model to derive the Marine Strategy Framework Directive indicator 1.5 (habitat area), assessing a descriptor of biological diversity. The research object encapsulates a scientific workflow, the input dataset, provenance information about the execution of the workflow, the output dataset, ancillary documentation such as images and presentations, and information regarding the

[4] http://purl.org/wf4ever/ro
[5] http://purl.org/wf4ever/wfdesc
[6] http://purl.org/wf4ever/wfprov
[7] http://purl.org/wf4ever/roevo
[8] Respectively, http://openarchives.org/ore and https://www.w3.org/ns/oa
[9] http://sandbox.rohub.org/rodl/ROs/SeaMonitoring01/

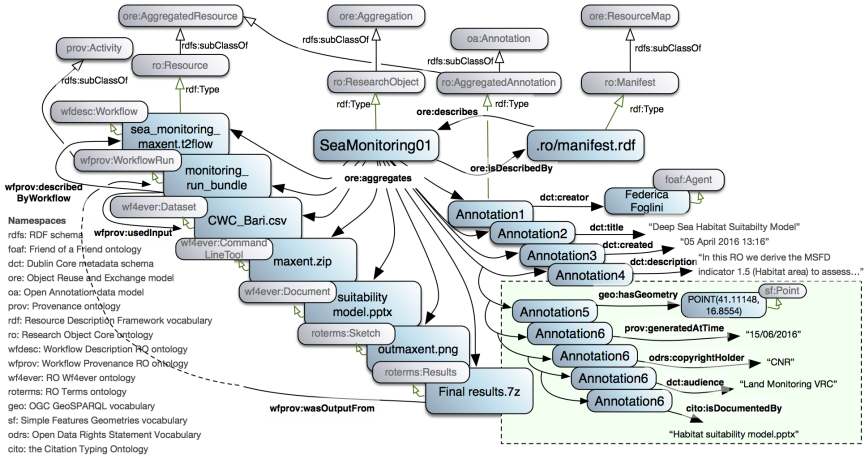


Figure 3: Simplified view of the research object containing a habitat suitability model (earth science specific metadata in the dashed rectangle).

author, plus metadata about the research object evolution and quality checks.

## 5. Research Object Model - Earth Science Extensions

In this paper we focus on scientific communities in Earth Science disciplines including sea monitoring, volcanology and biodiversity, that use earth observation data for different purposes. Such communities are represented by the following institutions.

- **Institute of Marine Science** (CNR-ISMAR)[10].
- **Geohazard Supersites and Natural Laboratories** (GSNL)[11], represented by the Italian National Institute of Geophysics and Volcanology (INGV).
- **National Ecological Observatory Network** (NEON)[12].

All these communities pursue FAIR practices for collaboration, sharing and reuse of scientific knowledge, even before actual publication of their work in conferences or journals. Two additional organizations focused on earth observation took part in our study, equally contributing requirements for the extension of the research object model and producing exemplary research objects: The UK Natural Hazards Partnership (NHP)[13], and the European Union Satellite Centre (SatCen)[14]. However, while the former three are focused on scientific research missions (and therefore fall in the scope of this paper), the last two serve operational purposes, providing earth observation services to a limited set of stakeholders and security agencies.

The research object model was developed initially in the context of experimental disciplines like genomics and astrophysics [16], where scientific workflows play a central role to enable reproducibility. However, though that is also a relevant aspect for Earth Science communities, these are more focused on observations, e.g. involving the analysis of time series satellite data, rather than experimentation. Therefore we carried out a gap analysis to identify the necessary updates to be implemented in the model. In doing so, we used three main channels [17]:

- **A requirements questionnaire** with 14 questions related to the intended use of research objects that was distributed to each of the four organizations.
- **A survey** addressed to the broader Earth Science community containing a subset of the above questionary, distributed ammong the participants of the Research Data Alliance RDA 9th Plenary Meeting[15].
- **Two Research Object Hackathons**, where 50+ users in total from the four organizations received training on research objects methods and tools and started modeling their own exemplars. In the first hackathon, delegates from other scientific domains like Astrophysics[16] also participated, sharing their experiences with research objects.

The analysis of the surveys and the hackathons revealed five main areas where the gap between the coverage provided by the research object model and the needs of earth scientists were significant: geospatial information, time-period coverage, intellectual property rights, data access policies, and general-purpose information. In some cases, such information was not covered at all by the previous version of the research object model (geographic, time, data access policies), and in other cases it was not covered with sufficient detail as required by the earth scientists (intellectual property rights). The main additions to the model are summarized below (details available in this technical report

[10] http://www.ismar.cnr.it
[11] http://supersites.earthobservations.org
[12] https://www.neonscience.org/
[13] http://www.naturalhazardspartnership.org.uk
[14] https://www.satcen.europa.eu

[15] https://www.rd-alliance.org/plenaries/rda-ninth-plenary-meeting-barcelona
[16] http://www.iaa.es

[18]) and illustrated in Figure 3 (see the annotations, and prefixes indicating the vocabularies used to model the new information, enclosed in the lower-right dashed rectangle).

- **Geospatial**, the coordinates of the region relevant for the research object and the observation it represents.
- **Time-period:** time span covered in the observation.
- **Intellectual property rights**, including copyright holder, copyright starting year, type of license and attribution.
- **Data access policy**, i.e. the access level and policies under which the research object can be accessed.
- **General metadata**, including the main scientific discipline of the research object, the size and format of the resources aggregated by the research object, the date when the research object was released, its digital object identifier (DOI), the status according to the research object lifecycle, and its target community.

The executable resources covered by the model have also been extended to cover not only scientific workflows but also other types of processes, such as web services, scripts, command line tools and dedicated software frequently used in Earth Sciences. Earth scientist also requested new types of research objects according to the kind of the aggregated resources. We extended the research object types to characterize not only workflow-centric research objects, but also data-centric and service-centric, as well as documentation and bibliographic research objects. Finally, the research object lifecycle was extended with a new status (forked), which characterizes a new branch of the research object derived from the main one.

While some of these changes were considered important for the overall research object community and were incorporated in the research object model[17], other updates were specific to Earth Sciences. Therefore we created a new branch in the code repository of the research object model containing all the new metadata elicited in our analysis[18].

### *5.1. Lifecycle Management Extensions*

The lifecycle refers to the different stages that a scientific research (and its associated research object) transitions, from hypothesis generation to publication and archival. In previous versions of the research object lifecycle [5], research objects could be *Live* (mutable research objects related to on-going research processes), *Snapshot* (immutable research objects derived from live research objects, that are ready to release intermediate results), and *Archived* (immutable research objects with final research results, where the research process has been completed). However, the creation of snapshots and archived research objects was limited to the authors of the particular research object, and hence other authors aiming to reuse intermediate results should wait until such snapshot was created. To cope with this limitation, and inspired in Open Source Software development practices, we introduced a *Fork* action[19] for public, live research objects. Forking a research object means to create a copy of the research object that could be used for testing new ideas without affecting the original research object, or start a new research process based on the forked research object, contributing to speed up research.

Another fundamental aspect that the original lifecycle lacked was the provisioning of DOIs for research objects. DOIs are an important tool to encourage scientists to change their current way of work to a one based on research objects since they can see the benefits of releasing intermediate results that will be properly credited. DOIs are aligned with the FAIR principles: i) they contribute to the findability of research data and methods, since they are persistent and searchable through a public DOI registry, and ii) they are dereferenceable, meaning that, through a single click, the user will be redirected to a landing page with the main metadata of the research object. Therefore we extended the lifecycle and associated infrastructure in ROHub[20] so that a DOI is automatically generated when a snapshot or and archived research object is released.

## 6. Extracting Content-based Research Object Metadata through NLP

The reuse of research objects depends to a large extent on their associated metadata. Metadata is key for scientists to evaluate if a given research object produced by someone else is suitable for their own needs, as a whole or partially. Similarly, it is also critical for computer systems, like search engines and recommenders, to automatically collect potentially relevant information through machine-readable annotations.

The research object model supports the generation of metadata enabling research object description from different viewpoints, including lifecycle information (status, evolution, quality checks, authors), resource types (document, workflow, dataset), and information derived from the actual content of such resources, like the specific research areas or the location of the investigation. It can also contain human annotations in titles, labels, descriptions, hypotheses, conclusions and comments. Amongst the different types of metadata, the latter is probably the most descriptive, accurate and valuable in order to obtain a deeper insight on the research since it deals with knowledge directly from the field. However, it formalization requires human involvement and tends to be neglected or embedded in unstructured documents of various formats, like technical reports, presentations or scientific papers. Despite its importance we found that content metadata is scarce for a large number of research objects. From a random

[17] https://github.com/ResearchObject/specifications/issues/13

[18] https://github.com/wf4ever/ro/tree/earth-science

[19] https://help.github.com/articles/fork-a-repo/

[20] ROHub is a node of DataCite and an authorized DOI provider.

sample of 2,500 research objects in ROHub only 800 have such basic content metadata as a descriptive title, with an average character count of 38. In addition, research object descriptions have a typical length of 138 characters, as concise as a Tweet.

### 6.1. Semantic Enrichment

To alleviate the scarceness of content descriptive annotations and to structure them beyond plain text, we propose to automatically enrich research objects with semantic metadata extracted from human-generated content in the research object, enhancing human and machine readability thus contributing to enable FAIR research and in line with related efforts like the Concept Web Alliance [19]. The resulting annotations are structured as semantic markup based on a knowledge graph [20] and included as annotations following the research object model. The enrichment process, depicted in Figure 7, comprises three main stages: the extraction of text from resources in the research object, the semantic analysis of such text, and the actual generation of semantic metadata.

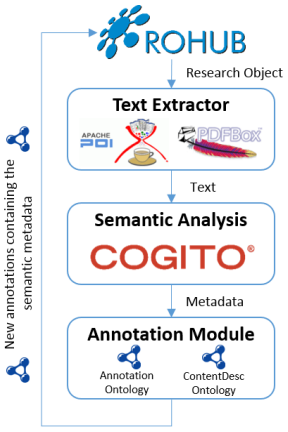


Figure 4: Semantic enrichment process

#### 6.1.1. Text Extraction

The enrichment process starts by gathering all the text available within research object resources and human annotations. We process resources in plain text, Microsoft Word and Powerpoint, and Adobe PDF formats, tagged as any of the following types[21]: Title (*dcterms:Title*), Description (*dcterms:Description*), Document (*wf4ever:Document*), BibliographicResource (*dcterms:BibliographicResource*), Conclusions (*roterms:Conclusions*), Hypothesis (*roterms:Hypothesis*), ResearchQuestion (*roterms:ResearchQuestion*), and Paper (*roterms:Paper*). We use open source tools to process PDF and Microsoft formats, such as apache PDFBOX and POI

[21] Resource type is assigned upon research object modeling in ROHub.

### 6.2. Semantic analysis

Research object enrichment builds on the semantic analysis of text[21], supported by tools such as DBpedia Spotlight[22], which uses Wikipedia articles as senses to annotate the text, or GATE[23], for ontology-based text annotation. Note that this paper focuses on the benefits of semantically annotating research object content beyond the actual tool producing such annotations. So, we will not compare the different alternatives avaliable. In this case we used Expert System's commercial platform Cogito[22] for convenience but could have chosen a different option. Rather than trying to cover the whole spectrum of metadata specified by the research object model, we focus on a more limited set of annotations supported by Cogito, that describe textual content at the domain level as follows:

- **Main Concepts** most frequently mentioned in a document. A concept groups words with the same meaning. E.g., *reservoir, artificial lake, man-made lake* are used to refer to *a lake used to store water for community use.*
- **Main Domains**: Fields of knowledge in which the main concepts are commonly used, e.g. *Hidrology* for the words in the former case.
- **Main Lemmas**: The cannonical form of the most frequent words in the text, e.g., *reservoir, artificial lake, and man-made lake.* A lemma can have different meanings and be associated to more than one concept, e.g. *reservoir* can also refer to *a person, animal, plant or substance in which an infectious agent normally lives and multiplies.*
- **Main Compound Terms**: Most frequent noun phrases[23], a group of words in a sentence that together behave as a noun. E.g., *water reservoir* or *hydrochemical element.*
- **Main Named Entities**: Most frequently mentioned named entities, i.e. People, Organizations and Places. E.g., *the black sea* is a place, *UN* is an organization, and *Elizabeth Mary* is a person.

Cogito is built on a knowledge graph (Sensigrafo), where concepts (syncons) are represented as groups of lemmas with the same meaning. Syncons are interconnected through semantic and linguistic relations, like hyperonymy, hyponymy and other properties. The English standard Sensigrafo we used in this work contains 301,582 syncons, 401,028 lemmas and 80+ relation types that yield about 2.8 million links. Among other purposes, Cogito leverages the knowledge contained in Sensigrafo to disambiguate the meaning of a word by recognizing its context.

#### 6.2.1. Annotation Generation

At the final stage we add the annotations produced by Cogito as research object metadata, following the annotation ontology, which is the standard way to annotate

[22] http://www.expertsystem.com/cogito

[23] http://dictionary.cambridge.org/dictionary/english/noun-phrase

```
@base: <.../LandMonitoring_Change_Detecting> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix cdesc: <https://w3id.org/contentdesc/> .

<.../ROs/LandMonitoring_Change_Detecting_Step>
  dc:subject <subject/1302006390>,<subject/280343272>,
             <subject/734754489>,<subject/1557562560>,
             <subject/1852089416>,<subject/79018874> .

<subject/1557562560> a "cdesc/Concept" ;
      skos:prefLabel "Segmentation and Reassembly" .

<subject/1852089416> a "cdesc/Concept" ;
      skos:prefLabel "Monitoring" .

<subject/79018874> a "cdesc/Domain" ;
      skos:prefLabel "Geology" .

<subject/280343272> a "cdesc/Domain" ;
      skos:prefLabel "Graphic" .

<subject/734754489> a "cdesc/Expression" ;
      skos:prefLabel "image processing algorithm" .

<subject/1302006390> a "cdesc/Expression" ;
      skos:prefLabel "exploitation of image archive" .
```

Listing 1: Example of semantic annotations

resources in the research object model, and the Content-Desc vocabulary (see `https://w3id.org/contentdesc`), which we developed to explicitly link these annotations to the semantics identified by Cogito. We have integrated the semantic enrichment service in ROHub as a nightly daemon, and a collection of semantically enriched research objects is available at `http://everest.expertsystemlab.com/browse`, including a search engine built on Solr[24].

*6.2.2. Semantic Enrichment Example*

The research object *Land Monitoring Change Detecting Step*[25] contains a workflow for change detection analysis and includes textual documents describing the hypotheses and conclusions of the analysis. The code excerpt in listing 1 shows the turtle[26] serialization of the semantic annotations added to the research object that were extracted from the textual content.

In this example the semantic enrichment added six pieces of metadata stating that the research object content, as defined by the *dc:subject* predicate, mainly refers to concepts (*cdesc/Concept*) "Monitoring" and "Segmentation and Reassembly", which fit in the "Geology" and "Graphic" domains (*cdesc/Domain*). Two of the most frequent compound terms or expressions (*cdesc/Expression*) are "exploitation of the image archive" and "image processing algorithm". Since the research object actually aims at detecting changes in a region by analysing satellite images and applying different image processing algorithms, the resulting metadata provides a rather accurate summary.

*6.2.3. Assessing the Relevance of the Semantic Metadata*

We asked members of the organizations participating in our study to answer a questionnaire regarding the new metadata added to the research objects. The objective was to assess the relevance of the annotation types (Domains, Concepts, Named Entities and Compound Terms) with which research objects are enriched against the research object content. In total, 10 researchers participated, who evaluated 19 research objects from their area of expertise and their annotations.

The analysis of the results [24] showed that domains and compound terms in general are perceived as relevant to the research object content, while concepts are also relevant but to a lesser extent, and named entities were not found useful by most of the evaluators. Domains are identified by aggregating the domains of all the concepts inferred from the text. Since we are reporting the most frequent domains in the text, erroneously identified domains are left in the long tail of the domain distribution. Compound terms, in turn, explicitly appear as expressions in the text, hence the high relevance perceived by the participants.

The results showed evidence that automatically produced semantic metadata brings about a positive enrichment of research object descriptions. They also suggest that dedicated user interfaces enabling users to act as curators of the annotations generated may be needed, since a fully automated solution is not feasible yet, given the state of the art in word sense disambiguation. However, we confirmed that a standard, out of the box version of Cogito can produce sufficiently good results for many of the target types of metadata, whose accuracy would be significantly improved, particularly for named entity recognition, with an extended version of Sensigrafo including additional Earth Science knowlege.

*6.3. Research Object Quality*

Research objects with high quality metadata are more likely to be reused than low quality ones, and in the long term such quality could experience changes, for example when some input file (e.g, an annotation file) becomes unavailable, degrading the overall quality of the research object and introducing decay. Inspired in wet lab practices checklists [25] were proposed as the main tool to assess the quality of research objects through their lifecycle [26]. These checklists are made up of statements that specify the required metadata a research object must contain.

A checklist contains the requirements that a research object must fulfill for a given purpose. It is not realistic to have a single set of criteria that fits all situations, i.e. the required metadata when reviewing an experiment differs from that involved in workflow execution. A requirement is a condition about the research object metadata and can be defined as mandatory, desirable, or optional. Requirements are validated through rules that describes how the requirement has to be tested. The most common type of rules are queries over the research object metadata to check for the existence of a particular piece of metadata.

Checklists collect the necessary information to calculate quality metrics about the completeness, stability and reliability of research objects[26]. Completeness measures

[24] `http://lucene.apache.org/solr`

[25] `http://sandbox.rohub.org/rodl/ROs/LandMonitoring_Change_Detecting/`

[26] https://www.w3.org/TR/turtle

the extent to which a research object satisfies a number of requirements specified in a checklist, stability measures the degree to which the research object completeness remains unchanged, and reliability combines both previous metrics to provide a unique value indicating to what extent the research object is complete and how stable it has been historically. These metrics are visualized in ROHub via an interactive chart displayed after clicking the RO monitoring tool link in the quality tab.

The hackathons allowed earth scientists to acquire experience with the research object model, create their own research objects and become aware of related benefits for their daily work. Scientists actually proposed specific new types of research objects to encapsulate mainly information regarding scientific workflows, data products, research products, and bibliographic information, which required to design different checklists to asses their quality [27]:

- **Basic**: This checklist addresses the minimum metadata required for a research object such as title, description, author, and access level. The rest of checklists presented below extend the basic checklist.
- **Workflow**: This checklist is intended for research objects built with a scientific workflow at the core. It tests metadata such as workflow definition, workflow execution, input and output data (including format and size), and workflow documentation.
- **Data Product**: This checklist addresses research objects containing mainly data sets. It checks metadata such as the purpose of the data, editor, copyright owner, access level, data format and size.
- **Research Product**, recommended for research objects dedicated to the analysis of data processing outcomes. It tests metadata such as the purpose, process implementation and input and output data.
- **Bibliographic**: This checklist is intended for research objects containing mainly bibliographic information such as bibligraphic references or documents that are a relevant to a specific topic. It tests metadata such as the copyright holder, the purpose and access level and the existence of at least one resource of type Bibliographic resource.

These checklists has been developed and made available in the Earth Science branch[27] in the research object github repository, and can be applied in ROHub to any research object in the Earth Science Domain.

## 7. Leveraging Research Object Metadata for Search and Recommendation

The research object metadata and text extracted from its payload can be leveraged by information retrieval tools that makes them visible to other researchers, thus improving their likelihood to be reused. Mainstream search engines are an important component since they reach a large number of users. ROHub allows web crawlers from Google and Bing indexing the research objects.

In addition, ROHub provides is own faceted search engine that uses the lifecycle metadata, the user-genetared metadata, and the content metadata generated by the semantic enrichment to ease the browsing of the research object collection. Facets allow the user to filter the collection by selecting specific values in properties (representing the facets) related to the research object (e.g., creator, or creation date). Some of these properties have values linked to a structured knowledge in the form of reference vocabulary (or ontology), such as *research area*, *type of research object*, *state of the life cycle*. Ontologies provides semantics to the property values, and enable semantic inference (e.g., a research object with research area astronomy, is also about space science).

Basic information about research objects is provided to external services through public search engine interface. It is implemented using OpenSearch specification `http://www.opensearch.org/` which makes it easily adopted by different clients and frameworks. ROHub's OpenSearch interface supports full text search for keyword based scenarios. In order to support finding research objects relevant to specific geographic region a spatial search extension was implemented. It allows usage of spatial intersection queries and returns georss elements `http://www.georss.org` in the output document.

Search engines are one of the tools of information retrieval, but not the only ones. Recommender Systems, on the other hand, support exploratory processes and search by example that could help researchers to find research works related to their own. In the following we describe a new recommender system that we developed benefiting of text within research objects and the metadata generated by the semantic enrichment.

### *7.1. Recommender System*

A recommender system[28] supports exploration when users do not know exactly what to search but have a partial knowledge of e.g. desired characteristics and related examples. Our recommender is content-based[29], i.e. user interests are expressed as a collection of research objects and matched against other research objects based on their content. This leverages the research object social dimension through forms of interaction among researchers such as research object coauthoring and citation.

We implemented a new recommender[28] based on the results of the experiments reported below, which exploits the metadata generated by the research object semantic enrichment process. The user interface built

[27] `https://github.com/wf4ever/ro/tree/earth-science/checklists`

[28] API at `http://everest.expertsystemlab.com/home/recommendation-api.html`

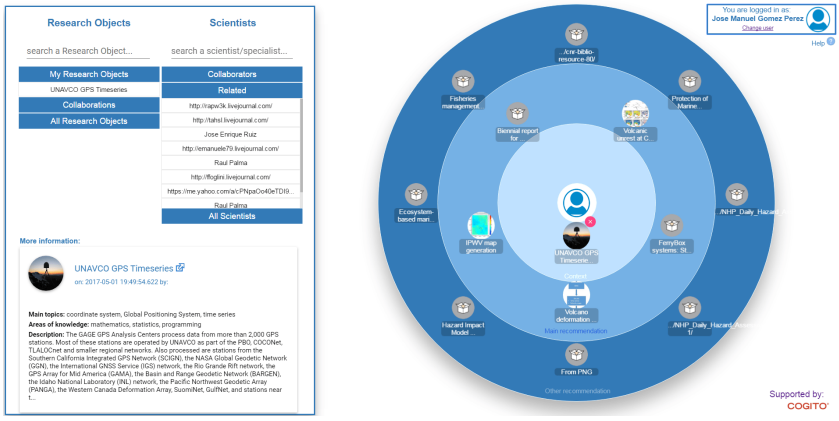


Figure 5: Collaboration Spheres: Recommender system user interface.

on top of it is shown in Figure 5. The system is accessible from `http://everest.expertsystemlab.com/spheres/index.html` and from ROHub (menu *Discover*).

The user interface follows a visual metaphor designed to facilitate research object sharing and reuse through goal-driven exploration of potentially large collections of research objects. It consists of a navigation panel and information card about the selected research object or scientist on the left-hand side, a set of concentric spheres on the right-hand side, and an authentication box and help option on the upper-right corner. Upon user authentication, the system produces personalized recommendations based on the collection of research objects (s)he authored. Through the navigation panel, the user can search for research objects or community members to be added to the recommendation context. The panel segments the collection of research objects in three subsets in decreasing order of proximity: the research objects authored by the user, those authored by collaborators, i.e. contributors to his or her research objects and the rest. Similarly for community members: collaborators, scientists related topic-wise and others.

The spheres component serves as a container for both the recommendation context and the recommendation results. Visually, the user is at the center of the spheres. The first sphere around it is an interactive area where the user can drag and drop up to three research objects, scientists (which, processing-wise, act as a proxy to their research objects), or a combination of both from the navigation panel in order to modify the recommendation context. The second and third concentric spheres display the recommendation results. The recommender assigns a score to each resulting research object, indicating its similarity with the recommendation context, which is used to sort the results. The higher the score, the closer to the center.

The usability and user satisfaction of the approach was assessed previously in [30]. Evaluators answered 50 questions[29] aimed at evaluating usability, user satisfaction, perceived usefulness and perceived ease of use. Average usability was 3.95 in a scale of 1 to 5, user satisfaction was 5.61 (1-7), and usefulness and ease of use scored 5.82 in the same scale.

### 7.2. Research Object Similarity

Research object recommendation builds on a notion of similarity between research objects in the collection and the ones included in the recommendation context. To calculate this similarity we use the traditional vector space model[31], whereby documents (i.e. research objects) and interests are mapped to vectors in a multimensional space where they can be compared using the cosine function as an indicator of similarity between them. Each dimension in this space is weighted according to a predefined weighting scheme[32] and corresponds to a keyword (or other kind of metadata) in the vocabulary that is used in the research object collection.

We carried out different experiments to better characterize the similarity measure, with different feature sets used to represent the research objects in the vector space model. The alternatives involved both the keywords extracted from the textual content in the research objects and the semantic metadata generated by the semantic enrichment process. We used the standard TF-IDF[30] as our weighting scheme. Note that the number of research objects in the Earth Science domain is still limited in ROHub since the community is just adopting the paradigm. Therefore we resorted to Wikipedia, where there is a good coverage of articles on Earth Science topics. The belonging of such articles to the domain can be easily determined through the categories assigned to them by the editors.

### 7.3. Experimental Setup

To generate the evaluation dataset we traversed the Wikipedia category graph starting in the Earth Science category[31], drilled down three levels in the subcategories, and

[29] Questions available at `https://sites.google.com/site/spheresquestionnaire/`

[30] TF-IDF stands for Term Frequency-Inverse Document Frequency.

[31] `https://en.wikipedia.org/wiki/Category:Earth_sciences`

collected all the articles annotated with these categories. We used DBpedia[32], the structured version of Wikipedia, to easily traverse the category graph. In total we harvested 27019 articles that were annotated with 1210 categories. We use such categories as indicators of similarity between articles. For each article we extracted the article title and textual content, discarding all the Wikipedia markup language tags, tables, references, image captions, and infoboxes. Then we created a research object for each article and proceeded to semantically enrich them.

To evaluate the similarity measure we use precision at k, a commonly used evaluation metric of ranked results in information retrieval[33]. In our case, precision measures the fraction of research objects identified by the similarity measure that are actually similar to the reference research object. Precision at k is computed on the subset of similar research objects until the k position of the ranked list of similar research objects. We repeated the experiments 10 times and report average precision (p) at 1, 5, 10 and 20.

### 7.4. Experiment 1

In the first experiment we calculated the similarity between a reference research object and the rest in the dataset. From our dataset we selected categories with at least 40 research objects, and randomly selected 10% of research objects in these categories. In total we assessed the similarity results regarding 2214 research objects under 250 categories. In addition to research objects in the same category, we used a relaxed definition of similarity where we considered as similar research objects also those in neighbor categories, i.e. subsumer (parent), siblings, and children categories. For example, the neighbor categories of *Marine Biology* are the subsumer *Oceanography*, the sibling *Marine Geology*, and the children *Marine Botany*, and *Cetology*. This similarity definition also indicates the variety of related research objects identified by the similarity measure, a desired property in recommender systems.

Table 3: Similarity Evaluation for one document

| Similarty based on | p@1 | p@5 | p@10 | p@15 | p@20 |
|---|---|---|---|---|---|
| **Similarity evaluated on same category** | | | | | |
| Concepts and text | 0,571 | 0,493 | 0,448 | 0,420 | 0,398 |
| Sem. metadata no NE and text | 0,565 | 0,490 | 0,445 | 0,417 | 0,396 |
| Sem. metadata and text | 0,569 | 0,490 | 0,445 | 0,417 | 0,396 |
| Concepts and NE and Text | 0,567 | 0,487 | 0,444 | 0,416 | 0,395 |
| Text (content+title) | 0,568 | 0,490 | 0,445 | 0,417 | 0,394 |
| Sem. metadata no NE | 0,480 | 0,415 | 0,378 | 0,355 | 0,339 |
| Sem. metadata | 0,481 | 0,412 | 0,373 | 0,350 | 0,335 |
| Concepts | 0,456 | 0,385 | 0,352 | 0,330 | 0,313 |
| Concepts and NE | 0,456 | 0,384 | 0,347 | 0,324 | 0,307 |
| **Similarity evaluated on neighbor categories** | | | | | |
| Concepts and text | 0,717 | 0,656 | 0,621 | 0,598 | 0,580 |
| Sem. metadata no NE and text | 0,718 | 0,654 | 0,620 | 0,597 | 0,579 |
| Text (content+title) | 0,718 | 0,657 | 0,620 | 0,597 | 0,578 |
| Concepts and NE and Text | 0,718 | 0,654 | 0,617 | 0,594 | 0,576 |
| Sem. metadata and text | 0,718 | 0,654 | 0,617 | 0,594 | 0,575 |
| Sem. metadata no NE | 0,643 | 0,590 | 0,559 | 0,538 | 0,523 |
| Sem. metadata | 0,639 | 0,578 | 0,548 | 0,527 | 0,513 |
| Concepts | 0,613 | 0,559 | 0,529 | 0,507 | 0,491 |
| Concepts and NE | 0,608 | 0,547 | 0,513 | 0,491 | 0,475 |

The experiment results are shown in Table 3, with the different approaches sorted in decreasing order by p@20. The best approach in both versions of the experiment was the combination of main concepts (top 10) generated by the semantic enrichment and textual content of the research object (concepts and text), followed by the combination of all the semantic metadata except named entities and textual content (semantic metadata no NE and text). In general, the combination of semantic metadata plus text seems to produce better results than semantic metadata alone. One interesting observation is that using only semantic metadata the precision values, albeit smaller, are close to other approaches using it in combination with text content. This supports our claim that automatically generated semantic metadata can alleviate the lack of user-generated metadata like research object title or description. Finally, although precision can still be improved, the similarity values evaluated on neighbor categories are promising.

### 7.5. Experiment 2

While the first experiment addressed one-to-one similarity-based recommendation, the second experiment aims at evaluating the similarity measure when the recommendation context includes the combined attributes of more than one research object. From the dataset, we randomly selected 1000 pairs of research objects where each pair was not annotated under the same category and the path between the categories in the category graph does not include the Earth Science category (since this would make the two resources barely related).

We use the category graph to determine the similarity between research objects by identifying the path connecting the categories of each of the two reference research objects, with the categories in such path as a similarity indicator. For example, if one of the reference research objects falls in the category *Oceanography* and the other one in the category *Marine Botany* we consider as similar research objects those falling in these categories plus the category *Marine Biology* since there exists the path *Oceanography* ⇒ *Marine Biology* ⇒ *Marine Botany*, where "⇒" means hasSubcategory.

We relaxed this definition by considering as similar objects those annotated with a category falling in the subtree whose root is the least common subsummer LCS [34] of the categories associated with the reference research objects. The LCS[33] is defined as the most specific common ancestor of two concepts found in a given ontology, and in our case it represents the semantic commonalities of the pair of categories. For example, the LCS of *Marine Biology* and *Ocean Exploration* is *Oceanography*. Similarly to experiment 1 this relaxed definition of similarity is aimed as an indicator of the variety of related research objects that the similarity measure generates. The experiment results are reported

[32] http://dbpedia.org

[33] http://www.igi-global.com/dictionary/least-common-subsumer-lcs/41765

in Table 4, where the different approaches are sorted in decreasing order by p@20.

Results, in table 4 show that using text information alone is the best approach when two research objects are used as the basis to obtain similar research objects. Nevertheless, the use of semantic metadata and text does not seem to harm, to a large extent, the precision of the similarity measure. In this experiment we also validated that the use of the semantic metadata without text produces, although smaller, similar results to the ones that we obtain when we have textual descriptions. The precision values of the similarity metric based on the LCS subtree are a good indicator of the usefulness of the metric in the recommender system when there are more than one research object in the recommendation context.

Table 4: Similarity Evaluation for context with two documents

| Similarty based on | p@1 | p@5 | p@10 | p@15 | p@20 |
|---|---|---|---|---|---|
| **Similarity evaluated on categories in the path** | | | | | |
| Text (content+title) | 0,577 | 0,492 | 0,445 | 0,417 | 0,406 |
| Sem. metadata no NE and text | 0,567 | 0,490 | 0,441 | 0,413 | 0,403 |
| Concepts and text | 0,571 | 0,489 | 0,442 | 0,412 | 0,401 |
| Sem. metadata and text | 0,563 | 0,485 | 0,439 | 0,410 | 0,399 |
| Concepts and NE and Text | 0,560 | 0,482 | 0,438 | 0,408 | 0,397 |
| Sem. metadata | 0,458 | 0,388 | 0,347 | 0,321 | 0,309 |
| Sem. metadata no NE | 0,448 | 0,387 | 0,343 | 0,321 | 0,308 |
| Concepts | 0,411 | 0,355 | 0,321 | 0,299 | 0,287 |
| Concepts and NE | 0,416 | 0,353 | 0,313 | 0,291 | 0,281 |
| **Similarity evaluated on categories in LCS subtree** | | | | | |
| Text (content+title) | 0,740 | 0,677 | 0,643 | 0,626 | 0,618 |
| Sem. metadata no NE and text | 0,732 | 0,677 | 0,641 | 0,623 | 0,616 |
| Concepts and text | 0,736 | 0,678 | 0,641 | 0,621 | 0,613 |
| Sem. metadata and text | 0,725 | 0,674 | 0,637 | 0,618 | 0,610 |
| Concepts and NE and Text | 0,724 | 0,673 | 0,636 | 0,615 | 0,607 |
| Sem. metadata no NE | 0,657 | 0,605 | 0,573 | 0,555 | 0,543 |
| Sem. metadata | 0,655 | 0,600 | 0,571 | 0,546 | 0,539 |
| Concepts | 0,617 | 0,583 | 0,549 | 0,530 | 0,520 |
| Concepts and NE | 0,614 | 0,576 | 0,535 | 0,515 | 0,506 |

## 8. Earth science interfaces for research objects

Enhancing traditional research practices with FAIR-enabled capabilities based on research objects requires specialized user interfaces that integrate the governance capabilities provided by research objects with existing tools already used by Earth scientist in their daily work. In doing so, we need to keep a delicate balance, pushing the boundaries of what is now possible with the current tools (i.e. adding new functionalities) while maintaining the familiarity with current interfaces and user experience.

In this section, we illustrate how this challenge has been addressed for different communities of scientists with specific needs and goals. The user interfaces and applications selected to that purpose include: the ROHub portal, the main front end for domain-independent research object lifecycle management sitting on top of the RO API; a Virtual Research Environment for vertical communities of scientists, in disciplines like sea monitoring and volcanology; and domain-specific applications dealing with time series data in the ecology and biodiversity domain.

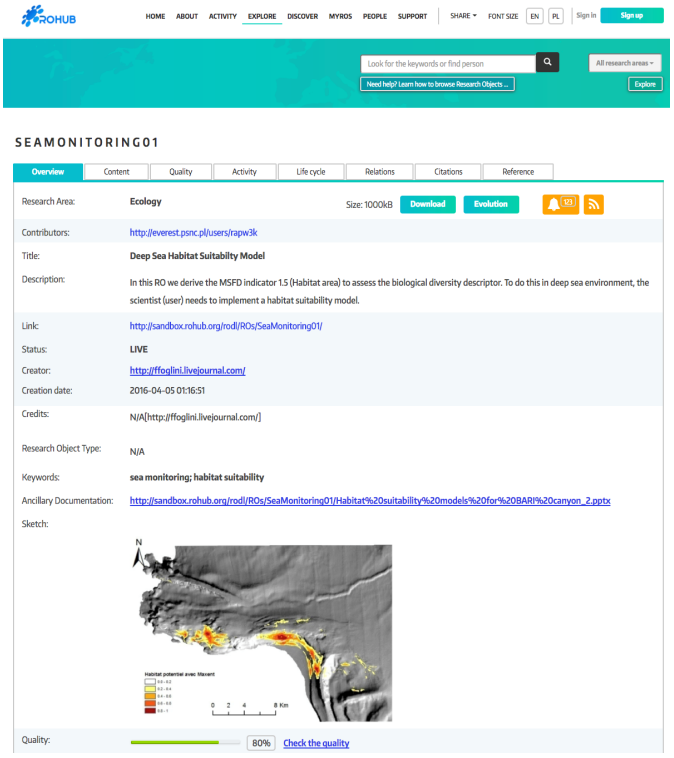


Figure 6: ROHub Portal

### 8.1. ROHub Portal

The ROHub portal is the generic front-end for research object management that provides an advanced, life cycle management-oriented, tool exposing the full set of research object management capabilities to scientists. It is intended for users who are already familiar with research objects, or who would like to analyze and manage research objects in at a finer grain of detail. Hence, it provides great flexibility and access to all possible operations at a granular level. In contrast, Virtual Research Community (VRC) portals for example (see section 8.2), provide scientists with access to composite custom-built operations at a higher level of abstraction. So, while in ROHub portal, the user may need to perform multiple individual operations to build a research object (create, annotate, add resources, etc.), the VRC portals encapsulate all these operations in a single, custom-built process.

The portal integrates and provides access to different research object services, including the core services provided by ROHub back-end for their creation, storage, access and maintenance, the management of their lifecycle, and their preservation, as well as added-value services like notification, transformation of workflows into research objects, quality and stability assessment, metadata enrichment, rating and exploratory search.

### 8.2. Community-Oriented Virtual Research Portals

Earth Science needs to address a variety of challenges. Among them, climate change is probably the most known

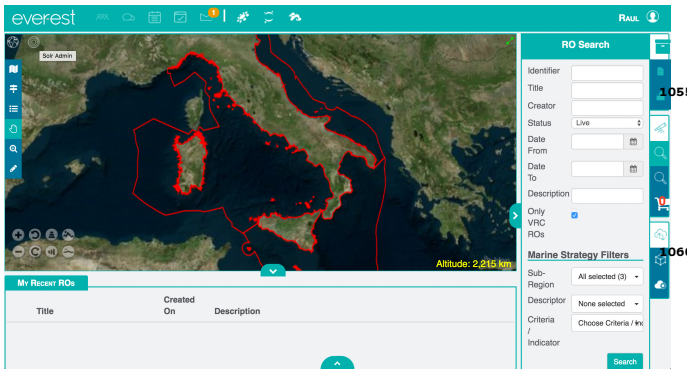


Figure 7: Sea Monitoring VRC Portal

topic because of its direct link to the increase of the average global temperature, but many others exist, including marine litter, air pollution, flooding and volcanic eruptions. This implies an increasing demand of data and information management capabilities to provide evidence, understand causes and monitor effects. The EVER-EST[34] virtual research environment (VRE) provides the different communities of earth scientists with virtual research community (VRC) portals offering custom services and tools targeted to ease work in community specific tasks. To support collaborative research across institutional and discipline boundaries, the VRE and VRC online portals use research objects to draw together research data, models, analysis tools and workflows as well as to manage and preserve the full research cycle. These interfaces abstract the research object vocabulary and details from the user, providing custom-built access to the core research object management capabilities in a simple and transparent manner. Currently there are four VRC portals - Land Monitoring[35], Natural Hazards[36], Sea Monitoring[37] and GeoHazards Supersites[38] - which can be accessed from the VRE, each pre-configured with the associated domain-specific data and services.

The VRC portals design reflects the UX shared among the Earth science communities and provides interactive tools to manage the full research in a 3D virtual globe, the most natural playground for an Earth Scientist to perform his/her activity. The toolbar on the right hand side (see figure 7) is the gamepad that collects [...to contninue from ] enables features related to research objects and other tools that are commonly used by earth scientists:

- **Research object management,** including basic research object functionalities, e.g. create, edit, annotate, etc, research object lifecycle management, metadata management or resource management.
- **Search:** Provides a search box to define search criteria for earth observation datasets and research objects.
- **Workflow discovery:** Enables the discovery and execution of scientific workflows by a generic workflow manager, e.g. Taverna server.
- **Virtual Machines:** Provides earth scientists with access to existing cloud resources, i.e. virtual machines, while enabling VRC administrators to manage them.
- **Web Processing Services (WPS):** Enables the discovery and execution of one of the available WPS.

### *8.3. Time Series Data Analysis in Ecology and Biodiversity*

Nowadays measuring the causes and effects of environmental change and how ecosystems are affected is a main concern for society and researchers. Scientists working on this problem often need to deal with data from different providers each of one serving the data they are specialized on. Scientists need to compare slices of time series data of different sensors and systems, keeping track of the provenance information that enable others to reproduce the experiments and reuse the results.

To support scientist interested in ecological processes we have developed and interactive web-based prototype application [39] (see Figure 8) that integrates time series from UNAVCO[40] and National Ecological Observatory Network (NEON)[41] sensors, and produces workflow-centric research objects. The UNAVCO stations record GPS positions while sensors in NEON towers provide multiple types of data, e.g., wind speed, humidity, etc., at different time resolutions. Users can plot and download time series data by selecting the station, sensor type, and time range.

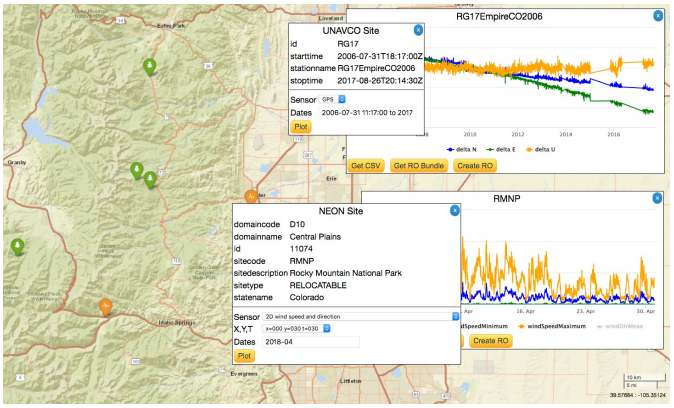


Figure 8: Web application to view UNAVCO and NEON time series.

Time series from UNAVCO and NEON are accessible from REST services. Since UNAVCO and NEON provide data in different formats, a workflow was developed in the Kepler Scientific Workflow System [35] to perform the REST queries and convert the results into GeoCSV [36]. A Kepler workflow consists of executable components, called "actors", linked together based on data dependencies to form

[34] http://vre.ever-est.eu
[35] http://vre.ever-est.eu/landmonitoring/
[36] http://vre.ever-est.eu/naturalhazards/
[37] http://vre.ever-est.eu/naturalhazards/
[38] http://vre.ever-est.eu/supersites/

[39] https://firemap.sdsc.edu/savi/map.html
[40] https://www.unavco.org/
[41] https://www.neonscience.org/

an overall application. The workflow for this application includes the actor to perform REST queries, and the R actor to convert data into GeoCSV.

After selecting time series from one or more sensors, a research object may be created to encapsulate the data and process used to create it. The research object includes a GeoCSV file containing the time series along with the instance of the Kepler workflow, which contains the parameters used to create the GeoCSV such as sensor location and time range. This workflow may be re-executed to produce the same time series data. The research object may either downloaded or shared on ROHub.

## 9. Community Adoption

We are still in early stages of the process to build a FAIR community of earth scientists that leverage research objects for the management, sharing and publication of their research and/or operational work on a normal basis. Nonetheless, the infrastructure is solid and we count with a considerable international community of early adopters, fundamentally distributed over Europe and the USA but also with some participation from Australia.

The different user interfaces built on top of the ROHub infrastructure presented in Section 8 have encouraged community members to move their work practices to those inspired by research objects and the FAIR principles. As a matter of fact, our early adopters are already producing and exploiting high quality research objects in both manual and automatic ways. And these research objects, as described in previous sections, are enabling these communities the adoption of the FAIR principles. That is, they are modeled based on interoperable ontologies, described through rich and expressive metadata, citable in scholarly communications, visible and discoverable from the Web and through recommendation systems, and ultimately, reusable.

Yet, to better understand the use made of the infrastructure by our community of scientists, to obtain a deeper insight and to facilitate the sustainability and continued growth of the community, we have implemented a number of mechanisms to monitor and measure our performance. In this section, we provide an account of the current progress stemming from quantitative data and related indicators.

### *9.1. Featured Research Objects*

Our early adopters increasingly use research objects and the associated infrastructure as part of their daily activities. After gaining a good understanding of the research object paradigm and the supporting technologies, key members of the community created a set of representative research objects for their area. We refer to the resulting research objects as Golden Exemplar Research Objects (GERO)[42] These are particularly curated and representative research objects that allow demonstrating the feasibility and utility of research objects to manage and share data, models and results of the daily work in Earth Science. Next, we select some of these golden exemplars from two of these communities, to further illustrate this approach:

***Sea Monitoring***

- **Detection of trends in the evolution of invasive jellyfish distribution**, a workflow-centric research object that produces explicit geographical information concerning the evolution and distribution of alien species based on Jellyfish sightings.
- **Digitalization of historical Venice lagoon maps**, a data-centric research object with information on natural environmental and anthropogenic changes.
- **Deep Sea Habitat Suitability Model**, a workflow-centric research object to derive the Marine Strategy Framework Directive MSFD indicator 1.5 to assess the biodiversity descriptor.

***Geoscience Research***

- **IPWV on Iceland**, a workflow-centric research object that automatizes the generation of a map of the precipitable water content on Iceland by using MODIS satellite data.
- **Volcano Source Modeling (VSM)**, a bibliographic research object containing all reports from March 2018 describing the weekly volcanic activity of Mt Etna from the multi-parametric monitoring stations.
- **UNAVCO GPS Position Timeseries**, a workflow-centric research object encapsulating a kepler workflow that calls a GPS position timeseries webservice provided by UNAVCO, processes the stream of data, and plots the north, east, and vertical offsets relative to a reference position

In addition to these manually crafted, high-quality research objects, we also generated through an automatic process over 500 bibliographic research objects[43] (AGROs - Automatically Generated Research Objects) exposing gray literature periodically released by these institutions, and bibliographic references of interest for the community.

### *9.2. Key Performance Indicators*

We have defined a set of key performance indicators (KPIs), consisting of measurable values, that allow us to: i) assess the success regarding the community adoption of research objects and related technologies; ii) estimate the extent to which this work is contributing to improve the currently limited compliance with the FAIR principles in Earth Science communities; and iii) to identify and analyze usage trends. For each of these KPIs, we defined a target for a six-month period, and started collecting the actual values monthly since the 2nd quarter of this year. The targets

[42] http://everest.expertsystemlab.com/#GoldenExemplars

[43] http://everest.expertsystemlab.com/#Generated

were defined with the feedback of key community members regarding their experiences and expectations about research objects and their daily work. So, each month we compare the measured values with the targets to assess the progress and to draw conclusions.

The KPIs are measurable via the ROHub platform, which integrates multiple added-value services and serves different client applications (see Section 8). Table (5) presents the KPIs, with the target values for the period (to the end of the 3rd quarter), and the last measured values (May 2018).

As we can observe from the table, we have already reached a few targets, including number of GEROs, number of AGROs and percentage of research object views. Reaching the targets in the number of golden and automatically generated research objects is a good indicator related to community adoption. But more importantly, having such significant number of research objects is an improvement in the FAIR level of these communities. In particular, now over 3500 data and other research artifacts are FAIR enabled via almost 750 research objects (see Section 3). In fact, reaching the target in the percentage of views is an evidence of that, i.e., resources are findable and accessible (see first two rows in Table 2).

Table 5 also shows that some KPIs are still below the target. However, in most cases the values measured are not so far from the targets, and we believe there is still enough time to reach the targets for the first period (end of 3rd quarter).For example, the number of resources managed by Earth Science communities through research objects is 64% below the target, while the average quality of research objects is only between 2 and 17% below. Note that quality related measurements take into account conditions like whether or not the data and associated research are well described (with rich, machine-readable metadata) or that resources are accessible, all of them key factors in terms of compliance with the FAIR principles. The fact that quality measures are almost aligned with the target values is a good indicator, showing evidence of convergence towards FAIR among the communities.

Nonetheless, indicators of reuse (research objects downloads and forks) are still far from the target and we have increased our efforts in analyzing how to raise such values. For instance, a better understanding is needed about how to encourage earth scientists to increase the share of work they do by reusing or repurposing existing results rather than by carrying out their research from scratch. Limited reuse values also indicates the need to provide earth scientists with means to simplify such tasks, lowering the technical entry barrier. Tooling support to enable proper credit to previous work, i.e. through persistent identifiers and enforcing automatic citation, is also key in this regard. Although such mechanisms are already available in ROHub (e.g. release of research objects with DOIs, research object fork and automatic citation to the source), our analysis seems to indicate a lack of awareness about such functionalities.

Furthermore, we have recently implemented in ROHub mechanisms that on the one hand enable scientists to express a subjective notion of quality about particular research objects and on the other hand keep account of the social impact of a research object among the user communities. Although the amount of data available to this purpose is still limited, we observe a trend indicating a correlation between research object reuse and their popularity. Frequently reused research objects have better ratings and reviews, and are favorited more frequently. As part of our awareness work, such features are now making their way into the user communities. Follow up work in this direction includes mechanisms to highlight or rank scientists depending on the reputation they earned based on the impact (rates, likes, views), reuse (downloads, forks) and quality of their research objects.

Table 5: Key performance indicators: targets (September 2018) against measures (May 2018)

| **Key Performance Indicator (KPI)** | | **Target** | | **Measured** | |
|---|---|---|---|---|---|
| Number of research objects implemented in Earth Science | | GEROs | 8 | GEROs | 16 |
| | | AGROs | 500 | AGROs | 512 |
| | | Total | 1000 | Total | 748 |
| Number of Earth Science resources managed by the communities | | Total | 10000 | Total | 3563 |
| Average quality of Earth Science research objects | | GEROs | 95% | GEROs | 93% |
| | | AGROs | 90% | AGROs | 73% |
| | | Released | 85% | Released | 72% |
| Impact of Earth Science research objects | Views | GEROs | 100% | GEROs | 100% |
| | | AGROs | 40% | AGROs | 99% |
| | Downloads | GEROs | 80% | GEROs | 44% |
| | | AGROs | 25% | AGROs | 2% |
| | Forks | Total | 25% | Total | 1% |

### *9.3. Web analytics*

Another mechanism that was put in place to monitor and to get insights about the adoption of research objects and related technologies is the tracking and reporting of ROHub web traffic using Google Analytics. We started tracking the ROHub Web site since March 1st 2018, and have already collected enough information to discover some patterns. For instance, figure 9 depicts the number of users visiting ROHub per day, where we can observe multiple peaks. After analyzing these peaks, we see that many of them coincide with the dates of dissemination or demonstration events, which indicates interest from the target communities, e.g., GeoVol (latin american workshop on volcanology) 7th-9th March, or EGU (European Geosciences Union) 9th-12th April. It is worth noting that since the beginning of the track history (83 days including weekends) only one day did ROHub not get any visit: Sunday 1st April (Easter).

Regarding the number of users per country, the USA is in first position, with about 23% of the share (see Figure 10). Although we have engaged some Earth Science communities there, this was an interesting finding. The second country is Poland (where ROHub is developed), followed by Italy (where two other important Earth Science communities are located), Spain (where another Earth Science community and a key technical partner are located), and the UK (where another Earth Science community is located).

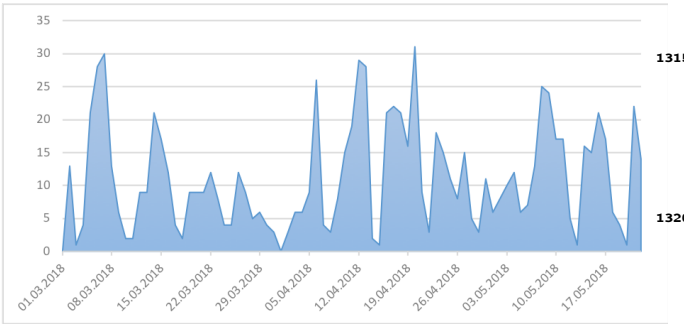


Figure 9: ROHub web traffic: users per day since March 2018

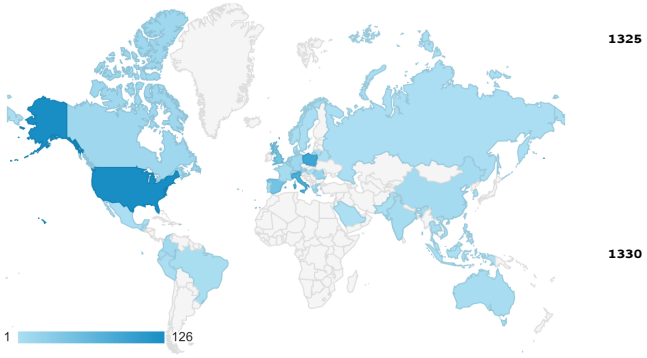


Figure 10: ROHub web traffic: users per country since March 2018

Another interesting discovery is that the busiest time of day is around noon, being 14:00 the busiest hour (based on the number of sessions), followed by 12:00, 11:00 and 15:00. This indicates that the busiest hour is right after lunch in Europe (CET time) and early morning in the United States (Eastern time), which seems to indicate that scientists actually access the platform as part of their daily routine. Regarding the busiest day of the week, we found no significant difference between working days, also indicating usage of the platform as part of the daily work activities.

## 10. Conclusions

In this paper we described the journey we went through to build a FAIR research environment for Earth Science around research objects. Transforming a data-intensive scientific community like this to use FAIR principles is a continuous and multidisciplinary effort that must be supported by methods, models and tools, while engaging early adopters from these communities to produce a critical mass of FAIR content that encourage their peers to adopt this new paradigm of work, leading to the establishment of a virtuous circle of FAIR data sharing and reuse.

Our work aimed at building upon the research object model a set of tools that ease the generation of research objects while increasing their likelihood to be reused by other researchers. Therefore our focus was on vocabulary extensions, automatic generation of metadata and quality assessment, search engines and recommender systems, digital object identifiers, and tailored user interfaces that incorporate earth science datasets, time-series data management and geolocalization. The key performance indicators to monitor the health of the research community of earth scientist working with research objects are in place. The challenge for the future is to enlarge the user community and leverage the experience gained with Earth Scientists to encourage other research communities to make the transition to a FAIR data interchange.

## Acknowledgements

We gratefully acknowledge EU Horizon 2020 for research infrastructures under grant EVER-EST-674907.

## References

[1] P. E. Bourne, T. W. Clark, R. Dale, A. de Waard, I. Herman, E. H. Hovy, D. Shotton, Improving The Future of Research Communications and e-Scholarship (Dagstuhl Perspectives Workshop 11331), Dagstuhl Manifestos 1 (1) (2012) 41–60. doi:10.4230/DagMan.1.1.41.
URL http://drops.dagstuhl.de/opus/volltexte/2012/3445

[2] A. M. Smith, D. S. Katz, K. E. a. Niemeyer, Software citation principles, PeerJ Computer Science 2 (2016) e86. doi:10.7717/peerj-cs.86.
URL https://doi.org/10.7717/peerj-cs.86

[3] H. Kitano, Artificial intelligence to win the nobel prize and beyond: Creating the engine for scientific discovery, AI Magazine 37 (1) (2016) 39–49.
URL http://www.aaai.org/ojs/index.php/aimagazine/article/view/2642

[4] S. Bechhofer, I. Buchan, D. D. Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, M. Gamble, D. Michaelides, S. Owen, D. Newman, S. Sufi, C. Goble, Why linked data is not enough for scientists, Future Generation Computer Systems 29 (2) (2013) 599 – 611, special section: Recent advances in e-Science. doi:https://doi.org/10.1016/j.future.2011.08.004.
URL http://www.sciencedirect.com/science/article/pii/S0167739X11001439

[5] K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, R. Palma, S. Bechhofer, E. Garcia-Cuesta, J. Gomez-Perez, G. Klyne, K. Page, M. Roos, J. Ruiz, S. Soiland-Reyes, L. Verdes-Montenegro, D. D. Roure, C. Goble, Workflow-centric research objects: A first class citizen in the scholarly discourse, in: 2nd Workshop on Semantic Publishing (SePublica), no. 903 in CEUR Workshop Proceedings, Aachen, 2012, pp. 1–12.
URL http://ceur-ws.org/Vol-903/paper-01.pdf

[6] J. Zhao, J. Gomez-Perez, K. Belhajjame, G. Klyne, E. García-Cuesta, A. Garrido, K. Hettne, M. Roos, D. D. Roure, C. Goble, Why workflows break - understanding and combating decay in taverna workflows., in: eScience, IEEE Computer Society, 2012, pp. 1–9.
URL http://dblp.uni-trier.de/db/conf/eScience/eScience2012.html#ZhaoGBKGGHRRG12

[7] A.-L. Barabási, Network theory–the emergence of the creative enterprise, Science 308 (5722) (2005) 639–641. arXiv:http://science.sciencemag.org/content/308/5722/639.full.pdf, doi:10.1126/science.1112554.
URL http://science.sciencemag.org/content/308/5722/639

[8] S. Crouch, N. C. Hong, S. Hettrick, M. Jackson, A. Pawlik, S. Sufi, L. Carr, D. De Roure, C. Goble, M. Parsons, The software sustainability institute: Changing research software attitudes and practices, Computing in Science and Engg. 15 (6)

(2013) 74–80. doi:10.1109/MCSE.2013.133.
URL https://doi.org/10.1109/MCSE.2013.133

[9] S. Hettrick, M. Antonioletti, L. Carr, N. Chue Hong, S. Crouch, D. De Roure, I. Emsley, C. Goble, A. Hay, D. Inupakutika, M. Jackson, A. Nenadic, T. Parkinson, M. I. Parsons, A. Pawlik, G. Peru, A. Proeme, J. Robinson, S. Sufi, Uk research software survey 2014 (Dec. 2014). doi:10.5281/zenodo.14809.
URL https://doi.org/10.5281/zenodo.14809

[10] M. Wilkinson, et al, The fair guiding principles for scientific data management and stewardship, Nature Scientific Data (160018).
URL http://www.nature.com/articles/sdata201618

[11] A. Dunning, M. de Smaele, J. Böhmer, Are the fair data principles fair?This version is provided as pre-print and is not peer reviewed practice paper. Please find the official and peer-reviewed publication via the INTERNATIONAL JOURNAL OF DIGITAL CURATION at the reference section on the right hand side. doi:10.5281/zenodo.321423.

[12] R. Palma, P. Hołubowicz, O. Corcho, J. Gomez-Perez, C. Mazurek, Rohub—a digital library of research objects supporting scientists towards reproducible science, in: Semantic Web Evaluation Challenge, Springer, 2014, pp. 77–82.

[13] C. Goble, D. De Roure, myexperiment: Social networking for workflow-using e-scientists, in: Proc. of the 2nd Workshop on Workflows in Support of Large-scale Science, WORKS '07, ACM, New York, NY, USA, 2007, pp. 1–2. doi:10.1145/1273360.1273361.
URL http://doi.acm.org/10.1145/1273360.1273361

[14] W3C, Rdf 1.1 primer: W3c working group note 24 june 2014, [Online; accessed 25-May-2017] (2014).
URL https://www.w3.org/TR/rdf11-primer/

[15] D. McGuinness, F. van Harmelen, Owl web ontology language overview: W3c recommendation 10 february 2004, [Online; accessed 25-May-2017] (2004).
URL http://www.w3.org/TR/owl-features/

[16] K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma, E. Mina, O. Corcho, J. Gomez-Perez, S. Bechhofer, G. Klyne, C. Goble, Using a suite of ontologies for preserving workflow-centric research objects, Web Semantics: Science, Services and Agents on the World Wide Web 32 (2015) 16 – 42. doi:https://doi.org/10.1016/j.websem.2015.01.003.
URL http://www.sciencedirect.com/science/article/pii/S1570826815000049

[17] J. Gomez-Perez, P. Alexopoulos, N. Garcia, R. Palma, D4.1 workflows and research objects in earth science — concepts and definitions, Tech. rep., European Virtual Environment for Research – Earth Science Themes (EVER-EST) (2016).

[18] J. Gomez-Perez, R. Palma, N. Garcia, D4.2 workflows and research objects models in earth science, Tech. rep., European Virtual Environment for Research – Earth Science Themes (EVER-EST) (2016).

[19] P. Groth, A. Gibson, J. Velterop, The anatomy of a nanopublication, Information Services & Use 30 (1-2) (2010) 51–56.

[20] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, F. Ciravegna, Semantic annotation for knowledge management: Requirements and a survey of the state of the art, Web Semantics: Science, Services and Agents on the WWW 4 (1) (2006) 14 – 28. doi:https://doi.org/10.1016/j.websem.2005.10.002.
URL http://www.sciencedirect.com/science/article/pii/S1570826805000338

[21] L. Reeve, H. Han, Survey of semantic annotation platforms, in: Proc. of the 2005 ACM Symposium on Applied Computing, SAC '05, ACM, New York, NY, USA, 2005, pp. 1634–1638. doi:10.1145/1066677.1067049.
URL http://doi.acm.org/10.1145/1066677.1067049

[22] P. Mendes, M. Jakob, A. Garcia-Silva, C. Bizer, Dbpedia spotlight: Shedding light on the web of documents, in: Proc. of the 7th Intl. Conference on Semantic Systems, I-Semantics '11, ACM, New York, NY, USA, 2011, pp. 1–8. doi:10.1145/2063518.2063519.
URL http://doi.acm.org/10.1145/2063518.2063519

[23] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. Greenwood, H. Saggion, J. Petrak, Y. Li, W. Peters, Text Processing with GATE (Version 6), 2011.
URL http://tinyurl.com/gatebook

[24] J. M. Gomez-Perez, R. Palma, A. Garcia-Silva, Towards a human-machine scientific partnership based on semantically rich research objects, in: 2017 IEEE 13th International Conference on e-Science (e-Science), 2017, pp. 266–275. doi:10.1109/eScience.2017.40.

[25] B. M. Hales, P. J. Pronovost, The checklist—a tool for error management and performance improvement, Journal of critical care 21 (3) (2006) 231–235.

[26] J. M. Gómez-Pérez, E. García-Cuesta, A. Garrido, J. E. Ruiz, J. Zhao, G. Klyne, When history matters-assessing reliability for the reuse of scientific workflows, in: International Semantic Web Conference, Springer, 2013, pp. 81–97.

[27] A. Garcia-Silva, J. Gomez-Perez, R. Palma, D4.3 design, implementation and deployment of research objects components for earth science phase 1, Tech. rep., European Virtual Environment for Research – Earth Science Themes (EVER-EST) (2016).

[28] P. Resnick, H. Varian, Recommender systems, Communications of the ACM 40 (3) (1997) 56–58.

[29] P. Lops, M. De Gemmis, G. Semeraro, Content-based recommender systems: State of the art and trends, in: Recommender systems handbook, Springer, 2011, pp. 73–105.

[30] M. Rico, J. M. Gómez-Pérez, R. Gonzalez, A. Garrido, Ó. Corcho, Collaboration spheres: a visual metaphor to share and reuse research objects, CoRR abs/1710.05604. arXiv:1710.05604.
URL http://arxiv.org/abs/1710.05604

[31] G. Salton, A. Wong, C. Yang, A vector space model for automatic indexing, Commun. ACM 18 (11) (1975) 613–620. doi:10.1145/361219.361220.
URL http://doi.acm.org/10.1145/361219.361220

[32] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing & Management 24 (5) (1988) 513 – 523. doi:http://dx.doi.org/10.1016/0306-4573(88)90021-0.
URL http://www.sciencedirect.com/science/article/pii/0306457388900210

[33] C. Manning, P. Raghavan, H. Schütze, et al., Introduction to information retrieval, Vol. 1, Cambridge university press Cambridge, 2008.

[34] Z. Wu, M. Palmer, Verbs semantics and lexical selection, in: Proc. of the 32nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, 1994, pp. 133–138.

[35] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, S. Mock, Kepler: an extensible system for design and execution of scientific workflows, in: Proceedings of 16th International Conference on Scientific and Statistical Database Management, IEEE, 2004, pp. 423–424.

[36] GeoWS Project, GeoCSV – Tabular text formatting for geoscience data (2015).
URL http://geows.ds.iris.edu/documents/GeoCSV.pdf