Towards a Human-Machine Scientific Partnership  
Based on Semantically Rich Research Objects

Jose Manuel Gomez-Perez  
Expert System  
Madrid, Spain  
jmgomez@expertsystem.com

Raul Palma  
Poznan Supercomputing and Networking Center  
Poznan, Poland  
rpalma@man.poznan.pl

Andres Garcia-Silva  
Expert System  
Madrid, Spain  
agarcia@expertsystem.com

Abstract—A research object is a single information unit encapsulating all the knowledge relevant to a particular scientific investigation, their associated metadata and the context where such resources were produced and came into play. Aimed at enhancing the preservation, reuse and scholarly communication of data-intensive science, research objects are both technical and social artifacts that represent a partnership between scientific communities and the computational support required in nowadays science. In this paper, we explore such partnership, identifying the lack of appropriate machine-readable metadata as one of its main inhibitors, and address the semantic enrichment of research objects as one key aspect towards its establishment. Focused on the specific needs of Earth Science communities, we propose extensions to research object representation models and present novel methods and tools to enrich research object metadata through automatic means. Finally, we validate the approach through the implementation of a recommender system that exploits the resulting metadata to facilitate research object discovery and reuse, enabling humans and machines to work together and accelerate the research life cycle.

I. INTRODUCTION

Much has been said in recent times about the expected impact of intelligent systems in many aspects of our lives. Today’s large amount of available data, produced at an increasing pace and in heterogeneous formats and modalities, has stimulated the development of means that extend human cognitive and decision-making capabilities, alleviating such burden and assisting our drivers, doctors, teachers and scientists, and sometimes even replacing them. In scientific disciplines like biomedical sciences, some even propose a new grand challenge for this kind of systems: to develop an AI that can make major scientific discoveries and that is eventually worthy of a Nobel Prize [1]. Though still far from realization, this scenario suggests the time is ripe for a shared partnership with machines, whereby humans can benefit from augmented reasoning and information management capabilities if machines are endowed with the necessary intelligence to assist with such tasks.

In data-intensive science we argue that research objects are one of the main enablers for such partnership between scientists and computers, with the potential to accelerate science. Conceptually speaking, a research object [2] is a container of scientific knowledge, a semantically rich aggregation of resources that brings together data, methods and people involved in a scientific investigation as a single information unit. Research objects encapsulate all the necessary information to preserve scientific work against potential decay [3] and can be shared, reused and cited in scholarly communications. As scholars move away from paper towards digital content, research objects have a key role to play in the way scientific results are communicated and validated by the communities, given the need for mechanisms that support the production of self-contained publications involving not only text but also data, methods and software implementations. Therefore, research objects serve a twofold purpose: they address technical challenges such as preservation, execution, interoperability and platform portability, but also enable social aspects of the scientific enterprise [4], like sharing and communicating results, checking their reproducibility, and giving credit to the authors.

A good part of the success of research objects as enablers of a prosperous partnership between scientists and intelligent systems enhancing scientific information management depends on the metadata describing research objects and their content. Without accurate, comprehensive, machine-readable metadata, the promise of automatic or at least assisted processing of scientific information as research objects seems unlikely.

Though critical for governance, discovery, sharing and reuse, research object metadata is typically generated manually by their authors, becoming labour-intensive and ultimately scarce. In addition, the complex structure of research objects, which can aggregate multimodal pieces of information such as documents, numerical datasets, pieces of programming code, workflows and the provenance of their executions, makes the description and retrieval of research objects a challenging task. Indeed, the research object model [5] does not impose any constraint to the types of content that may be included in these objects. Thus, research object metadata is mainly related to the structure and tends to concentrate at the container level without actually exploring the payload, which may involve such valuable sources of information as scientific papers, field notes, presentations and technical reports. Consequently, basic information tokens that can provide a summarized view of the research object at the domain level, including the main concepts related to the investigation, the scientific areas it addresses, relevant names of persons, places and organizations, and frequently used expressions, remain hidden to any automated means of discovery and search.

The resulting lack of visibility can reduce the potential outreach and diffusion of scientific outcomes, hindering research
object reuse by other teams of scientists and ultimately posing an obstacle for the incremental development of science. The limited availability of metadata is therefore one of the main obstacles for scientists to benefit from the adoption of research objects, but it can be alleviated by putting in place the necessary mechanisms to extract the required metadata from the research object content itself.

This paper describes the journey of introducing research objects in Earth Sciences, from the understanding of these communities in terms of representing, disseminating and reusing scientific knowledge to the required extensions of the research object representation formalisms and the associated technological support. The work presented 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. In this paper we focus on the application of research objects to four different communities in Earth Sciences, extending previous work in experimental disciplines like Genomics and Astrophysics, and validate the approach through the implementation of a research object recommendation system that leverages the metadata produced by the semantic enrichment mechanism.

**Paper structure**

Section II introduces the research object notion and the main building blocks of the research object model. Next, Section III describes the process followed to create research object awareness amongst the Earth Science communities, and how this resulted in the extension of the research object model and its adoption by such communities. In Section IV, we argue that the scarceness of metadata and more specifically of user-generated annotations about research object content is an important limiting factor for the dissemination and reuse of scientific knowledge as research objects, one of the main benefits expected by earth scientists. In this section we propose to address such limitation through the automatic generation of research object metadata extracted from content in textual form. Then in Section V we validate, through the development of a research object recommender system and user interface, how the metadata resulting from the automatic mechanism along with human-generated metadata enhances research object discoverability by humans and machines and potential reuse. Finally, in Section VI we present the conclusions of this work.

**II. RESEARCH OBJECTS**

In order to collaborate in a so-called partnership for enhanced science, scientists and assisting systems need to represent scientific knowledge in a form, rich with annotations, that makes it recognizable, processable, and exchangeable by both humans and machines. This artifact is what we call a research object, a semantically rich aggregation of resources that bundles together essential scientific information relating to an investigation [2]. This information is not limited merely to the data used and the methods employed to produce and analyse 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 [2] [6]. 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.org [7] is the reference platform for research object management, with myExperiment.org as its nearest precursor [8].

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, essential to interpret the scientific outcomes captured by a research object.

The research object model relies on the W3C Resource Description Framework RDF [9], a data model specifically designed for data interchange in the web, and the Web Ontology Language OWL [10], 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** (ro), describing the aggregation of resources in the research object, as well as the annotations made on those resources.
- **The Workflow Description Ontology** (wfdesc), meant as an upper ontology for more specific workflow definitions, and as a way to express abstract workflows.
- **The Workflow Execution Provenance Ontology** (wfprov), for the representation of provenance information generated by the execution of a scientific workflow.
- **The Research Object Evolution Ontology** (roevo), which describes research object lifecycle information.

Aggregation is supported through the use of the OAI-ORE vocabulary and annotation is supported through the use of the Web Annotation Ontology [1]. Moreover, 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.

1. [http://purl.org/wf4ever/ro](http://purl.org/wf4ever/ro)
2. [http://purl.org/wf4ever/wfdesc](http://purl.org/wf4ever/wfdesc)
3. [http://purl.org/wf4ever/wfprov](http://purl.org/wf4ever/wfprov)
5. Respectively, [http://openarchives.org/ore](http://openarchives.org/ore) and [https://www.w3.org/ns/oa](https://www.w3.org/ns/oa)
Figure 1 shows a graphical representation of an existing research object 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 author, plus metadata about the research object evolution and quality checks.

III. RESEARCH OBJECTS IN EARTH SCIENCE

Originally validated in experimental sciences [5], in this paper we expand on this and report our experience bringing the research object concept to observational disciplines, particularly Earth Sciences. We are piloting the adoption of the research object concept in four main Earth Science communities:

- **Sea Monitoring**, represented by the Italian Institute of Marine Science (CNR-ISMAR).\(^6\)
- **Natural Hazards**, through the UK Natural Hazards Partnership (NHP).\(^7\)
- **Land Monitoring**, represented by the European Union Satellite Centre (SatCen).\(^8\)
- **Geohazard Supersites and Natural Laboratories** (GSNL), represented by the Italian National Institute of Geophysics and Volcanology.\(^9\)

All these communities pursue the twofold goal of scientific knowledge long-term preservation and enhanced practices for collaboration, sharing and reuse, even before actual publication [11]. CNR-ISMAR seek to use research objects to share and reuse data and models that enable the assessment of Good Environmental Status (GES) descriptors and of measurable environmental targets for such descriptors. GSNL aim at leveraging research objects for better sharing and disseminating scientific knowledge, ensuring the reproducibility of results for validation. From a more operational viewpoint, NHP pursue more effective means to validate and share mathematical hazard impact simulations that remain accountable after the actual assessment, while SatCen are interested in applications to automatically detect changes in time series of satellite images.

**Extensions to the model**

The research object model was developed initially in the context of experimental disciplines like Genomics and Astrophysics [5], where scientific workflows play a central role to enable reproducibility. However, though these are also relevant aspects for Earth Sciences, observational disciplines are more focused on other aspects, involving mainly the analysis of time series data. 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 [11]:

- **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 among the participants of the Research Data Alliance RDA 9th Plenary Meeting.\(^11\)
- **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

---

\(^6\)http://sandbox.rohub.org/rodl/ROs/SeaMonitoring01/
\(^7\)http://www.ismar.cnr.it
\(^8\)http://www.naturalhazardspartnership.org.uk
\(^9\)https://www.satcen.europa.eu
\(^10\)http://supersites.earthobservations.org
\(^11\)https://www.rd-alliance.org/plenaries/rda-ninth-plenary-meeting-barcelona
other scientific domains like Astrophysics 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 Everest deliverable 4.2 [12]) and illustrated in Figure 1 (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**, the 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 submission date when the research object was released, its digital object identifier (DOI), the status according to the research object life cycle, and the main 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, and developed the associated checklists to assess their quality. 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 github repository, 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.

**Featured Research Objects**

Based on the experience gained on the research object model and its Earth Science extensions as well as supporting tools, our earth scientists started creating research objects containing data analysis models ready to be released amongst their peers. They produced the following so-called Golden Exemplar Research Objects to demonstrate the feasibility and utility of research objects to share data, models, and results that are representative of the daily work in Earth Science. In addition to these manually crafted research objects, we also produced a mechanism for the automatic aggregation as bibliographic research objects of scientific literature, including links to publications and grey literature, e.g. technical reports.

- **The Citizen science and jellyfish distribution**, computing MSFD indicators on jellyfish population from data gathered by a crowdsourcing app.
- **Trend analysis in the evolution of invasive jellyfish distribution**, which produces explicit geographical information concerning the evolution and distribution of alien species based on Jellyfish sightings.
- **Natural Hazards**
  - **Hazard Impact Model**, for surface water flooding simulation and early warning systems.
  - **Land Change Detection**, which aims at detecting anomalies in time series of satellite images acquired on specific locations and their correlation with social sensing sources and other geotagged information.
  - **Geohazard Supersites and Natural Laboratories**
    - **Volcano Source Modelling (VSM)**, modeling ground deformation from satellite InSAR (Interferometric Synthetic Aperture Radar) data observed at Campi Flegrei (Italy) during 2011-2013, in order to derive magmatic source characteristics.
    - **IPWV map generation**, which automates the generation of a map of integrated precipitable water vapor distribution over the Etna supersite, through satellite and GPS time series data.

**IV. SEMANTIC ENRICHMENT OF RESEARCH OBJECTS**

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 from 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,  

---

12. [http://www.iaa.es](http://www.iaa.es)
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 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.

To alleviate the scarceness of this kind of 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, also in line with related efforts like the Concept Web Alliance [13]. The resulting annotations are structured as semantic markup based on a knowledge graph [14] and included as annotations following the research object model. The enrichment process, depicted in Figure 2, 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.

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: 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.

Semantic analysis

Research object enrichment builds on the semantic analysis of text [15], supported by tools such as DBpedia Spotlight [16], which uses Wikipedia articles as senses to annotate the text, or GATE [17], 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 available. In this case we used Expert System’s commercial platform Cogito 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. Hydrology for the words in the former case.
- **Main Lemmas**: The canonical 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, 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 called Sensigrafo, where concepts (called 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 the context where it occurs.

---

17 Resource type is assigned upon research object modeling in ROHub.
18 See https://pdfbox.apache.org and https://poi.apache.org, respectively.
19 http://www.expertsysten.com/cogito
20 http://dictionary.cambridge.org/dictionary/english/noun-phrase
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 resources in the research object model, and the ContentDesc 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.

Enrichment Example

The research object Land Monitoring Change Detecting Step 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 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.

Assessing the Relevance of the Semantic Metadata

We asked members of the four 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. Figure 3 summarizes the results.

The analysis of the results shows 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.

Concepts were deemed less useful than expected, with only a slightly positive ratio. However, this is not entirely surprising since word sense disambiguation is still an open problem where state of the art tools such as [18] and [19] produce f-measure figures around 0.59 according to SemEval [20]. Plus, we (purposefully) used a standard version of Cogito, without extensions for the Earth Science domain. We also found participants sometimes felt confused when presented the main lemma of the concepts identified by Cogito, which not necessarily was the same as the actual word used in the text (e.g. soil vs. earth). As to the reason why named entities were rated as slightly non relevant, we found that, due to the lack of domain specialization, the system was confused when it came to disambiguate names. For example, the earth observation program Copernicus was confused with the astronomer Nicolaus Copernicus, simply because the former was not known to the system.

The results show 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 knowledge. Finally, an interesting finding relates to the cognitive gap between how concepts are referred to in the text and semantically equivalent terminological alternatives, and how such gap produces a (negative) effect in the perception of the user.

V. RECOMMENDER SYSTEM

We developed a research object recommender system with the goal of stimulating research object reuse within communities of scientists and to validate the impact of the automatically

---

21http://lucene.apache.org/solr
22http://sandbox.rohub.org/rodl/ROs/LandMonitoring_Change_Detecting/
23https://www.w3.org/TR/turtle
24F-measure is the harmonic mean of precision and recall.
generated metadata for enhanced research object discoverability by both humans and machines. A recommender system [21] supports exploration when users do not know exactly what to search but instead have partial knowledge of e.g. desired characteristics and related examples or community members. Our recommender is content-based [22], 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.

**User Interface**

For the earth science communities we implemented a new recommender\(^{25}\) 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 on top of it is shown in Figure 4. Currently accessible from http://everest.expertsysteclab.com/spheres/index.html it will soon be available in ROHub, too.

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 [23]. Evaluators answered 50 questions\(^{26}\) 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.

**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 [24], whereby documents (i.e. research objects) and interests are mapped to vectors in a multidimensional 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 [25] 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\(^{27}\) 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

\(^{25}\)API at http://everest.expertsysteclab.com/home/recommendation-api.html

\(^{26}\)Questions available at https://sites.google.com/site/spheresquestionnaire/

\(^{27}\)TF-IDF stands for Term Frequency-Inverse Document Frequency.
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.

Experimental Setup

To generate the evaluation dataset we traversed the Wikipedia category graph starting in the Earth Science category, drilled down three levels in the subcategories, and collected all the articles annotated with these categories. We used DBpedia, 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 [26]. 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 the average precision (p) at 1, 5, 10 and 20.

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.

The experiment results are shown in Table I, 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.

---

**Table I**

**Similarity Evaluation for One Document**

<table>
<thead>
<tr>
<th>Similarity evaluated on same category</th>
<th>p@1</th>
<th>p@5</th>
<th>p@10</th>
<th>p@15</th>
<th>p@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concepts and text</td>
<td>0.571</td>
<td>0.493</td>
<td>0.448</td>
<td>0.420</td>
<td>0.398</td>
</tr>
<tr>
<td>Semantic metadata no NE and text</td>
<td>0.565</td>
<td>0.490</td>
<td>0.445</td>
<td>0.417</td>
<td>0.396</td>
</tr>
<tr>
<td>Semantic metadata and text</td>
<td>0.569</td>
<td>0.490</td>
<td>0.445</td>
<td>0.417</td>
<td>0.396</td>
</tr>
<tr>
<td>Concepts and NE and Text</td>
<td>0.567</td>
<td>0.487</td>
<td>0.444</td>
<td>0.416</td>
<td>0.395</td>
</tr>
<tr>
<td>Text (content-title)</td>
<td>0.568</td>
<td>0.490</td>
<td>0.445</td>
<td>0.417</td>
<td>0.394</td>
</tr>
<tr>
<td>Semantic metadata no NE</td>
<td>0.480</td>
<td>0.415</td>
<td>0.378</td>
<td>0.355</td>
<td>0.339</td>
</tr>
<tr>
<td>Semantic metadata</td>
<td>0.481</td>
<td>0.412</td>
<td>0.373</td>
<td>0.350</td>
<td>0.335</td>
</tr>
<tr>
<td>Concepts</td>
<td>0.456</td>
<td>0.385</td>
<td>0.352</td>
<td>0.330</td>
<td>0.313</td>
</tr>
<tr>
<td>Concepts and NE</td>
<td>0.456</td>
<td>0.384</td>
<td>0.347</td>
<td>0.324</td>
<td>0.307</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Similarity evaluated on neighbor categories</th>
<th>p@1</th>
<th>p@5</th>
<th>p@10</th>
<th>p@15</th>
<th>p@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concepts and text</td>
<td>0.717</td>
<td>0.656</td>
<td>0.621</td>
<td>0.598</td>
<td>0.580</td>
</tr>
<tr>
<td>Semantic metadata no NE and text</td>
<td>0.718</td>
<td>0.654</td>
<td>0.620</td>
<td>0.597</td>
<td>0.579</td>
</tr>
<tr>
<td>Text (content-title)</td>
<td>0.718</td>
<td>0.657</td>
<td>0.620</td>
<td>0.597</td>
<td>0.578</td>
</tr>
<tr>
<td>Concepts and NE and Text</td>
<td>0.718</td>
<td>0.654</td>
<td>0.617</td>
<td>0.594</td>
<td>0.576</td>
</tr>
<tr>
<td>Semantic metadata and text</td>
<td>0.718</td>
<td>0.654</td>
<td>0.617</td>
<td>0.594</td>
<td>0.575</td>
</tr>
<tr>
<td>Semantic metadata no NE</td>
<td>0.643</td>
<td>0.590</td>
<td>0.559</td>
<td>0.538</td>
<td>0.523</td>
</tr>
<tr>
<td>Semantic metadata</td>
<td>0.639</td>
<td>0.578</td>
<td>0.548</td>
<td>0.527</td>
<td>0.513</td>
</tr>
<tr>
<td>Concepts</td>
<td>0.613</td>
<td>0.559</td>
<td>0.529</td>
<td>0.507</td>
<td>0.491</td>
</tr>
<tr>
<td>Concepts and NE</td>
<td>0.608</td>
<td>0.547</td>
<td>0.513</td>
<td>0.491</td>
<td>0.475</td>
</tr>
</tbody>
</table>

---

29http://dbpedia.org
(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.

**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* \(\Rightarrow\) *Marine Biology* \(\Rightarrow\) *Marine Botany*, where "\(\Rightarrow\)" 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 subsumer LCS [27] of the categories associated with the reference research objects. The LCS\(^{30}\) 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 in Table II, where the different approaches are sorted in decreasing order by p@20.

Results 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.

### VI. Conclusions

This paper describes our experience bringing research objects to a scientific domain like Earth Science in order to enable a social and technical foundation towards a prosperous partnership for scientific information management between scientists and assistive intelligent systems. Inspired by our previous work in experimental sciences, earth scientists saw in research objects an instrument to preserve, validate, share and reuse scientific data, models and result and to cross-fertilize ideas in observational disciplines. However, this vision requires rich metadata that successfully describes research object structure, content and goals. To this purpose, in collaboration with different communities in Earth Science, we understood the metadata that is important in the domain and extended the original research object model with means to represent information such as geographical coordinates, time-based data, copyright information, and access policies.

We investigated if the metadata required to guarantee that research objects are visible by scientists and systems, as a primordial step for reuse, is being generated by their authors and found out it is actually scarce. Thus, we propose to automatically enrich research object metadata by extracting it from their actual content in text form, relieving users from this manual effort and alleviating the lack of human-generated metadata. We also noticed a lack of tools to support (earth) scientists to effortlessly explore collections of research objects and implemented a recommender system that leverages the combination of automatic and human-generated annotations to calculate a similarity measure between a recommendation context and relevant elements of such collections. Through the recommender we validated the benefits of automatically

### Table II

<table>
<thead>
<tr>
<th>Similarity evaluated on categories in the path</th>
<th>Similarity evaluated on categories in LCS subtree</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Similarity based on</strong></td>
<td><strong>p@1</strong></td>
</tr>
<tr>
<td>Text (content-title)</td>
<td>0.577</td>
</tr>
<tr>
<td>Semantic metadata no NE and text</td>
<td>0.567</td>
</tr>
<tr>
<td>Concepts and text</td>
<td>0.571</td>
</tr>
<tr>
<td>Semantic metadata and text</td>
<td>0.563</td>
</tr>
<tr>
<td>Concepts and NE and Text</td>
<td>0.560</td>
</tr>
<tr>
<td>Semantic metadata</td>
<td>0.458</td>
</tr>
<tr>
<td>Semantic metadata no NE</td>
<td>0.448</td>
</tr>
<tr>
<td>Concepts</td>
<td>0.411</td>
</tr>
<tr>
<td>Concepts and NE</td>
<td>0.416</td>
</tr>
</tbody>
</table>

**Concepts and NE and Text**

<table>
<thead>
<tr>
<th><strong>Semantic metadata and text</strong></th>
<th><strong>Concepts and text</strong></th>
<th><strong>Semantic metadata and text</strong></th>
<th><strong>Concepts and NE and Text</strong></th>
<th><strong>Semantic metadata no NE</strong></th>
<th><strong>Semantic metadata no NE</strong></th>
<th><strong>Concepts</strong></th>
<th><strong>Concepts and NE</strong></th>
<th><strong>0.614 0.576 0.535 0.515 0.506</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Text (content-title)</td>
<td>0.740</td>
<td>0.677</td>
<td>0.643</td>
<td>0.626</td>
<td>0.618</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic metadata no NE and text</td>
<td>0.732</td>
<td>0.677</td>
<td>0.641</td>
<td>0.623</td>
<td>0.616</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concepts and text</td>
<td>0.736</td>
<td>0.676</td>
<td>0.641</td>
<td>0.621</td>
<td>0.613</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic metadata and text</td>
<td>0.725</td>
<td>0.674</td>
<td>0.637</td>
<td>0.618</td>
<td>0.610</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concepts and NE and Text</td>
<td>0.724</td>
<td>0.673</td>
<td>0.636</td>
<td>0.615</td>
<td>0.607</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic metadata no NE</td>
<td>0.657</td>
<td>0.605</td>
<td>0.573</td>
<td>0.555</td>
<td>0.543</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic metadata</td>
<td>0.655</td>
<td>0.600</td>
<td>0.571</td>
<td>0.546</td>
<td>0.539</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concepts</td>
<td>0.617</td>
<td>0.583</td>
<td>0.549</td>
<td>0.530</td>
<td>0.520</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concepts and NE</td>
<td>0.614</td>
<td>0.576</td>
<td>0.535</td>
<td>0.515</td>
<td>0.506</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^{30}\)http://www.igi-global.com/dictionary/least-common-subsumer-lcs/41765
generating research object metadata. Quality is in general close to human-generated metadata, showing evidence that we can effectively alleviate the lack of it through synthetic means. The future development of dedicated semantic annotation resources for the Earth Science domain will make this gap even smaller. We will also explore crowd-based approaches for vocabulary definition now being discussed by the community [28] and their complementarity with our approach.

This work is already paving the way for other scientific communities in experimental and earth disciplines to adopt research objects, like NEON, the flagship NSF project on Biodiversity. We have published a number of materials that serve this purpose, including research object exemplars, extensions and refinements of the research object model, and tools that enhance the exploration and reuse of research objects at http://everest.expertsystemlab.com. Ongoing work at advancing the use of research objects as means to support scholarly communications. Through our partnership with DataCite, it will soon be possible to assign DOIs to research objects in ROHub, enabling research object citation. A research object containing the supporting material, datasets, and relevant links related to this paper is available at http://sandbox.rohub.org/rodli/ROs/experiences-escience-2017/.

ACKNOWLEDGMENTS

We gratefully acknowledge funding from the EU Horizon 2020 program for research infrastructures under grant EVER-EST-674907. The authors would like to thank all the earth scientists who collaborated in this effort and Raul Ortega for his contributions.

REFERENCES


http://www.neonscience.org