



*Research Lifecycle Management technologies for
Earth Science Communities and Copernicus users in EOSC*

Deliverable D3.3

Design, implementation and deployment of text mining and enrichment services Phase 2

Grant agreement number	101017501
Start date of the project	Reliance
Duration of the project	30 months
Type of Action	Research and Innovation action
Coordinator	PSNC

Due date of delivery	30/09/2022
Actual date of delivery	05/10/2022
Work package	WP3
Type of deliverable	Report
Dissemination level	Public
Responsible	Expert.ai
Reviewer	Daniel Garijo and Esteban Gonzalez (UPM)
Version	2.0



This project has received funding from the European research infrastructures (including e-Infrastructures) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017501

List of authors, contributors and reviewers

Name	Role	Organization
Raul Ortega	Author	Expert.ai
Andres Garcia-Silva	Author	Expert.ai
Cristian Berrío	Author	Expert.ai
Jose Manuel Gomez-Perez	Author	Expert.ai
Daniel Garijo	Reviewer	UPM
Esteban Gonzalez	Reviewer	UPM

History of changes

Version	Date	Change	Authors	Organization
0.1	01/08/2022	Table of Contents added	Andres Garcia	Expert.ai
0.1.1	13/09/2022	Executive Summary	Andres Garcia	Expert.ai
0.1.2	19/09/2022	Chapter 3 added	Raul Ortega, Cristian Berrío	Expert.ai
0.1.4	19/09/2022	Chapter 4 added	Raul Ortega	Expert.ai
0.1.5	20/09/2022	Chapter 5, 6 and Conclusions	Raul Ortega, Andres Garcia	Expert.ai
0.1.6	22/09/2022	Internal review	Jose Manuel Gomez Perez	Expert.ai
0.9	29/09/2022	Final version after internal review	Andres Garcia	Expert.ai
1.0	29/09/2022	Final internal review and approval	Jose Manuel Gomez-Perez	Expert.ai

Glossary

Acronym	Explanation
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
FoR	Field Of Research
IPTC	International Press Telecommunications Council
NLP	Natural Language Processing
NL API	Natural Language API
LM	Language Model
OpenAIRE	Open Access Infrastructure for Research in Europe
RELISH	REliance daSHboard
REST	Representational State Transfer
ROHub	Research Object Hub
SQUAD	Stanford Question Answering Dataset

Table of Contents

1	<i>Executive Summary</i>	6
2	<i>Introduction</i>	7
2.1	<i>Scope</i>	7
2.2	<i>Structure</i>	8
3	<i>Extended analytics services in support of the scientific enterprise</i>	8
3.1	<i>Influence Networks: claim analysis</i>	8
3.2	<i>Challenges and Solutions Extractor</i>	10
3.3	<i>Novelty Score</i>	12
3.4	<i>Reading Comprehension: Question Generation</i>	14
3.5	<i>Enrichment Web Application</i>	15
4	<i>Semantic annotations and information extraction: Enrichment API</i>	15
4.1	<i>Document Enrichment</i>	15
4.2	<i>Research Object Enrichment</i>	17
4.3	<i>Document Enrichment Demo</i>	18
5	<i>Enhanced Content-based Retrieval</i>	22
5.1	<i>RELISH: RELiance daSHboard</i>	22
6	<i>Conclusions and future work</i>	23
7	<i>References</i>	24

List of Figures

Figure 1.	Claim Analysis pipeline	8
Figure 2	Novelty score service diagram	13
Figure 3.	Summary of the Document Enrichment workflow. In green the metadata available in the first stage. In blue the metadata added in this stage.	16
Figure 4.	Screenshot of the upper part of the Document Enrichment Demo	19
Figure 5.	Screenshot of the Topics section in the Document Enrichment Demo	20
Figure 6.	Screenshot of the Key Elements section in the Document Enrichment Demo	20
Figure 7.	Screenshot of the Entities section in the Document Enrichment Demo	21
Figure 8.	Screenshot of the Extended Analytics section in the Document Enrichment Demo	21
Figure 9.	Screenshot of RELISH	22
Figure 10.	Screenshot of the action bar from RELISH	23
Figure 11.	Screenshot of the Concepts panel from RELISH	23
Figure 12.	Screenshot of the lowermost part of RELISH with the filtered list of research objects	23

1 Executive Summary

This deliverable reports the work carried out during the second year of the project within work package 3. During this period the goal was increasing the robustness of the services deployed during the first year and extending the currently available text analytics and understanding services with the development of the Extended Analytic Services described in task T3.4. In the first year we deployed in EOSC an **enrichment service**¹ to generate structured metadata about the content of a research object or a document including the main concepts and phrases, plus the entities and their type (e.g., people, location, organization), and topical information from domains according to the Expert.ai linguistic knowledge graph. To demonstrate the enrichment service capabilities, we published an **enrichment web application**².

Most of the metadata produced by the first version of the enrichment service aims to increase the research object findability by providing structured information describing its content. We leveraged such metadata to feed a **semantic search engine**³ and a **recommendation service**⁴ for research objects that we also deployed in EOSC. The front-end of the recommendation service is the **Collaboration Spheres**⁵ web application where users can drag and drop research objects or their authors to the recommendation context and receive a recommendation of research objects based on that content. In the second year of Reliance, we move on from services supporting research object findability to services supporting research object comprehension and termed such services as the Extended Analytic Services. Here the goal is to assist researchers in the comprehension of the research work represented by a research object and how such research relates to other work in the scientific community, including other research objects and scientific publications. The new services include: i) the **challenge and solution extraction** service to extract the scientific challenges addressed by the work described in a research object and the corresponding solution or research proposal as long as this information is contained in the research object description, ii) the **Question Generation** service that produces questions from a research object description and highlights in the text the answer to such questions, iii) the **claim analysis** service, which identifies scientific claims in research objects and associates them with existing claims in the scientific literature, and iv) the **novelty scoring** service that assigns a numeric value between 0 (non-novel) and 1 (novel) based on the assumption that the novelty of the research proposed in a research object is inversely proportional to the amount of existing similar work. The extended analytics services require advanced text processing capabilities at the edge of the State of the Art in natural language processing and understanding, including machine-reading comprehension, natural language generation, text classification and semantic textual similarity. Even though we use state-of-the-art language models fine-tuned for the underlying task addressed by each service, the extended analytics services need to be considered as experimental since such tasks are still open research problems in NLP/U. To emphasize the maturity level difference between the production-ready enrichment and the new set of Extended Analytic services we deployed the latter as separate web services⁶.

In addition, we enhance the enrichment service by increasing the types and formats of the documents it supports. We add support to Jupyter Notebooks and GitHub repositories. We also add new metadata such as lemmas, main sentences, categories from taxonomies including Fields of Research, NASA Scope and Subject, and IPTC, as well as entity types like Data Cubes identifiers, and time references.

Finally, we have designed **RELISH** (REliance DaSHboard), a dashboard where ROHub administrators can visualize the metadata extracted by the text mining services to have a high-level overview of the

¹ <https://marketplace.eosc-portal.eu/services/enrichment-api>

² <https://reliance.expertcustomers.ai/enrichment/>

³ <https://marketplace.eosc-portal.eu/services/search-api>

⁴ <https://marketplace.eosc-portal.eu/services/recommendation-api>

⁵ <https://reliance.expertcustomers.ai/spheres>

⁶ Extended services are under the path: <https://reliance.expertcustomers.ai/extended/>

content of the research objects hosted in the platform. The dashboard is interactive and provides different widgets to visualize the aggregated information of the research object collection. We include charts to show the distribution of research objects according to the most representative metadata, including the categories of the taxonomies we use to classify them and tag-cloud visualizations among others to show concepts, lemmas, phrases, and entities.

2 Introduction

The objective of work package 3 is to provide EOSC and its user communities with text mining and analytics services that leverage the wealth of unstructured text data from different sources such as scholarly communications and technical reports, text descriptions associated to datasets and/or accessible as Data Cubes, and from documents contained in research objects.

To realize such objective in task 3.1 we first gather a corpus of text relevant to reliance user communities that was used to integrate the domain-specific vocabulary in the models and knowledge representation underlying the text mining services. Next in task 3.2 we develop a service to enrich semantically documents and research objects with structured metadata describing their content. To demonstrate the enrichment service capabilities, we develop an enrichment dashboard⁷ where users can choose a pre-defined text or paste a new one and see the metadata that our service generates.

In addition, in task 3.3 we leverage the new metadata added to the research object collection in ROHub to implement enhanced information retrieval tools. We develop a semantic search service, and a recommendation service with the Collaboration Spheres as its user interface. The enrichment service, the search service and the recommendation service are available in EOSC portal⁸.

To disseminate the text mining services developed in RELIANCE we publish an informative web page with documentation and links to the services in EOSC⁹. All the above work is reported in the first version of this deliverable.

In this second version of the document, we describe new services developed in task 3.4 that aim at enhancing the comprehension of research objects and their context in the scientific field. In contrast to the previous version of the enrichment service where the metadata is used to fuel information retrieval tools, the metadata generated by the new services is intended to ease the understanding of the research object content to researchers. In addition, we have developed the RELISH dashboard where it is possible to have a high-level overview of the research object collection in ROHub by showing their distribution in terms of the taxonomies and structured metadata used to enrich them. Finally, we enhance the enrichment service to generate new metadata and process markdown text in Jupyter Notebooks and GitHub repositories and notify errors in the service to ROHub.

2.1 Scope

The scope of this document is concerned with the services developed in task 3.4 and the modifications we have made to strengthen the services and demos developed in the other tasks. Concretely we present:

- New extended analytics services for: i) Challenge and Solution extraction, ii) Question Generation, iii) Claim Analysis, and iv) Novelty Score.
- A new error notification and callback mechanism between the enrichment service and ROHub.
- New document formats supported: Jupyter Notebooks and GitHub repositories.
- New types of extracted metadata: Data cube identifiers and new taxonomies supported by the enrichment service.
- Extension of the document enrichment web application with the new services and resources.
- A newly developed RELISH dashboard (RELIance daSHboard).

⁷ <https://reliance.expertcustomers.ai/enrichment/>

⁸ <https://marketplace.eosc-portal.eu/services?providers=262>

⁹ <https://reliance.expertcustomers.ai/>

2.2 Structure

In section 3 we present the Extended Analytic Services developed in task 3.4. Next, in section 4 we describe the features added to the enrichment service including new taxonomies supported in the scientific domain, and the error handling mechanism to notify ROHub when the enrichment fails. Finally, section **Error! Reference source not found.** describes the RELISH dashboard and section 6 presents the conclusions.

3 Extended analytics services in support of the scientific enterprise

The Extended analytic services are intended to increase the researcher's comprehension of research objects and its context in the research field. Services like the challenge and solution and the question generation service are useful to understand research objects. On the other hand, services like the novelty score and the claim analysis services help to identify a research object's related work and how such research object fits in the context of such work. In addition, in this section we include the description of the latest version of the enrichment web application where we integrate the new metadata added by the extended analytics services and the latest version of the enrichment service.

3.1 Influence Networks: claim analysis

An influence network depicts the relationships between scientific works and how they influence each other. A straight way to establish the influence relationships is to leverage citations from one paper to another since a citation is an indicator that the cited paper is important in the context of the current research. A more interesting and challenging approach is to establish the influence network at the level of scientific claims. That is, to relate a scientific claim detected in a source research work with a similar claim in another research work previously published.

We understand as a scientific claim a statement that can be found within the scientific literature. It can be an assertion about a specific scientific subject such as "The life cycle of ferns is characterized by two phases: gametophyte and sporophyte", and it should be verifiable using a reliable and contrasted source. Our goal is therefore to use claims to connect research objects and scientific publications. The claim analysis pipeline extracts, compares, and connects claims between research objects and scholarly communications in an external scientific repository. The input of this pipeline is the description of a research object from ROHub, and the output is one or more pairs of claims: one claim from the research object and another from the scientific repository.

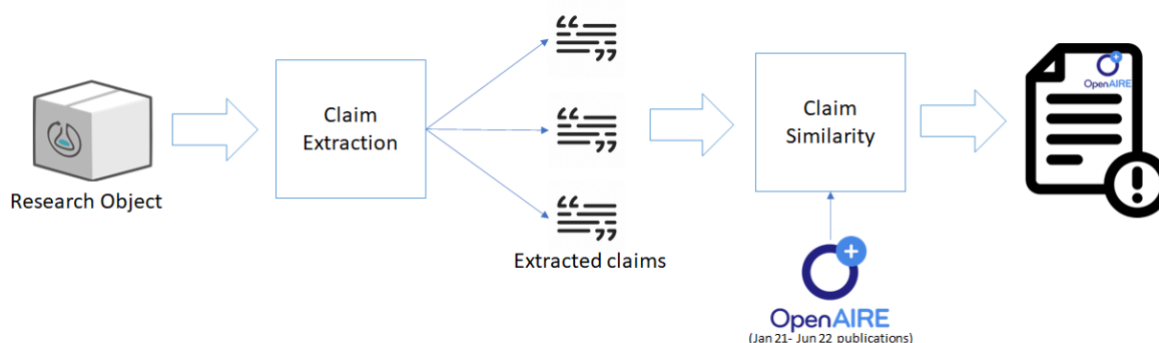


Figure 1. Claim Analysis pipeline

As shown in the Figure 1, the claim analysis pipeline consists of three main components:

- **Claim extraction module.** This module oversees the selection of sentences from the description of the research object that can be considered a claim. To do so, we fine-tune two pre-trained language models using the transformers architecture (Vaswani, et al., 2017) in a

binary classification task. The input of these models is a sentence, and the output whether the sentence is a claim or not. The finetuning of the models (RoBERTa Base (Liu, et al., 2019)) was performed over four datasets: CLEF2019¹⁰, Claim Data Buster¹¹, Poynter¹² and Cite-Worth (Wright & Augenstein, 2021). The first three datasets are general purpose datasets that are used to fine-tune the first model, which tries to extract general claims, such as statements that can be part of a publication or a journal article. The latter is a scientific-domain dataset, and it is used to fine-tune the second model, which tries to extract cite alike claims.

- **External scientific repository.** A repository of reliable articles that can function as a trustworthy dataset of scientific literature. We select OpenAIRE publications as the base to generate a first version of the external scientific repository, but it will be completed with more trustworthy papers during the rest of the project. OpenAIRE is a very interesting repository since it contains a plethora of open access articles as opposed to other repositories where articles are behind a paywall. We split the abstracts of the OpenAIRE articles in sentences to compare them with the claims extracted with the extraction module. We use ColBERT (Khattab & Zaharia, 2020) to efficiently retrieve similar claims from research objects to sentences in abstracts. ColBERT generates embeddings representations for the sentences that can be queried to obtain the most similar sentence given an input claim. We select the three most similar sentences to each claim and pass them to the next component of the pipeline.
- **Claim similarity module.** This module uses semantic similarity (Cer, Diab, Agirre, Lopez-Gazpio, & Specia, 2017) to compare a claim and a related sentence. The semantic similarity model is a RoBERTa base (Liu, et al., 2019) fine-tuned in a sentence similarity task with the STS-B dataset (Cer, Diab, Agirre, Lopez-Gazpio, & Specia, 2017). We choose RoBERTa base due to its good performance on the STS-B dataset where the model reaches a Pearson-Spearman correlation of 91.2. The output of this model is a score from 0 to 5. We select those pairs of claims that obtain at least a score of 3.5. This value, as mentioned in (Cer, Diab, Agirre, Lopez-Gazpio, & Specia, 2017), implies that the two sentences are at least roughly equivalent, but some valuable information can be missed or be different. Having roughly equivalent as the minimum similarity allows the system to locate related claims that might agree or disagree with each other.

The claim analysis service can help users to visualize the connections between research objects from ROHub and an external collection of publications, giving the users a contextualized view of their research.

We have deployed this service as a REST API that can be called with a piece of text as body and returns a JSON document with a summary of the claim analysis. The JSON document includes the score given by each model of the pipeline and the ID of the source from the external scientific repository. The service can be called by making a POST request to the URL: https://reliance.expertcustomers.ai/extended/claim_analysis. An example of a request can be found below:

```
POST /extended/claim_analysis HTTP/1.1
Host: reliance.expertcustomers.ai
Content-Type: text/plain
Content-Length: xx
```

Coronavirus disease 2019 (COVID-19), first reported in Wuhan, the capital of Hubei, China, has been associated to a novel coronavirus, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). In March 2020, the

¹⁰ <https://sites.google.com/view/clef2019-checkthat/task-1-check-worthiness>

¹¹ <https://idir.uta.edu/claimbuster/>

¹² <https://www.poynter.org/ifcn-covid-19-misinformation/>

World Health Organization declared the SARS-CoV-2 infection a global pandemic. Soon after, the number of cases soared dramatically, spreading across China and worldwide. Italy has had 12,462 confirmed cases according to the Italian National Institute of Health (ISS) as of March 11, and after the 'lockdown' of the entire territory, by May 4, 209,254 cases of COVID-19 and 26,892 associated deaths have been reported. We performed a review to describe, in particular, the origin and the diffusion of COVID-19 in Italy, underlying how the geographical circulation has been heterogeneous and the importance of pathophysiology in the involvement of cardiovascular and neurological clinical manifestations.

Next, we can see the JSON output of the service:

```
[
  {
    "claim": "In March 2020, the World Health Organization declared the SARS-CoV-2 infection a global pandemic",
    "verified_claim": [
      {
        "docno": "21422b462af0ee4705c2ea2e1bcl69ed_s0",
        "extraction_score": 28.12066078186035,
        "similarity_score": "0.8562994",
        "verified_claim": "The novel SARS-CoV-2 outbreak was declared as pandemic by the World Health Organization (WHO) on March 11, 2020.\n"
      },
      {
        "docno": "7562c69833eedef2af3977365ff88711_s1",
        "extraction_score": 27.13343620300293,
        "similarity_score": "0.80039823",
        "verified_claim": "This cluster quickly spread across the globe and led the World Health Organization (WHO) to declare severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) a pandemic on March 11, 2020.\n"
      },
      {
        "docno": "de1e305d3bc8629c48678cf9b6050b73_s0",
        "extraction_score": 27.083131790161133,
        "similarity_score": "0.70187604",
        "verified_claim": "Following the emergence of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) responsible for COVID-19 in December 2019 in Wuhan (China) and its spread to the rest of the world, the World Health Organization declared a global pandemic in March 2020.\n"
      }
    ]
  }
]
```

3.2 Challenges and Solutions Extractor

This service extracts the main challenge and the main solution given the title and the description of a research object. To achieve this, we fine-tuned a language model (RoBERTa (Liu, et al., 2019) base and large) to classify a sentence as challenge, solution, or none, using a dataset containing texts from an innovation management platform¹³ annotated with problems and solutions. The dataset¹⁴ was

¹³ <https://ideas.esa.int>

¹⁴ Due to license restrictions in the source data, we cannot publish the dataset for the time being

annotated by a group of 7 annotators, has a total of 300 texts, and a subset of 20 texts were annotated by all the annotators so that the inter-rater agreement could be calculated as quality metric of the annotation process. After users annotated the texts, the inter-annotator agreement was computed using Cohen's kappa and Fleiss' kappa metrics:

- Average Cohen's kappa at sentence level: 0.5133
- Fleiss' kappa at sentence level: 0.5108

The level of agreement in the previous metrics is sometimes regarded as moderate agreement, which is in an indicator of the dataset quality.

The dataset contains a total of 2339 sentences labelled as "none", in contrast to 306 and 384 sentences labelled as challenges and solutions, respectively. To balance the dataset, we reduced the number of "none" sentences to 385. We split the dataset of 1075 sentences, taking randomly the 80% for training (860), and 20% for testing (215). Once the models are trained with the training set, we evaluate them with the testing set. The evaluation results of the different trained models for the classification task are presented in Table 1, showing that the base model gives better results than the large model.

Table 1. Evaluation results of sentence classifiers in Problem, Solution or Neither

	PRECISION	RECALL	F1
baseline-random	0.320	0.330	0.316
RoBERTa base	0.696	0.684	0.683
RoBERTa large	0.659	0.661	0.659

The title and description are sent in JSON format to the challenge and solution extractor API available at <https://reliance.expertcustomers.ai/extended/csextractor> using a POST request as for example:

POST /extended/csextractor HTTP/1.1

Host: reliance.expertcustomers.ai

Content-Type: application/JSON

Content-Length: xx

```
{
  "title": "Further to the Left: Stress-Induced Increase of Spatial
Pseudoneglect During the COVID-19 Lockdown",
  "description": "Background The measures taken to contain the
coronavirus disease 2019 (COVID-19) pandemic, such as the lockdown in
Italy, do impact psychological health; yet less is known about their effect
on cognitive functioning..."
}
```

The challenge and solutions extractor API uses the fine-tuned RoBERTa model to classify all the sentences in the Research Object description, takes the sentences classified as solutions, and calculates their similarity with the title. The sentence with the highest similarity score is selected as the main solution, then the API takes the sentences classified as challenges, and calculates the similarity with the title, selecting the two most similar sentences. Finally, the closest sentence to the main solution is chosen as the main challenge. The response is returned in a JSON containing the main challenge and main solution, for example:

```
{
  "error": null,
  "results": {
```

```

    "challenges": [
      {
        "end": 217,
        "start": 0,
        "text": "Background The measures taken to contain the
coronavirus disease 2019 (COVID-19) pandemic, such as the lockdown in
Italy, do impact psychological health; yet, less is known about their
effect on cognitive functioning."
      }
    ],
    "solutions": [
      {
        "end": 710,
        "start": 467,
        "text": "The aim of the present study was to investigate
the possible effects and impact of the COVID-19 pandemic on spatial
cognition tasks, particularly those concerning spatial exploration, and the
physiological leftward bias known as pseudoneglect."
      }
    ]
  }
}

```

3.3 Novelty Score

This service calculates the novelty score of a research object. We define the novelty score of a research object r , given a collection of other research objects R , and a collection of publications P , with the following formula:

$$NoveltyScore(r, R, P) = 100 * (1 - \max(\text{similarity}(r, R), \text{similarity}(r, P)))$$

The rationale behind the formula is that, if a research object is similar to an existing research work (in the collection of research objects R or in the publications P), the novelty score would be very low and vice versa. The range of similarities is between 0 and 1, and consequently the range of the novelty score is between 0 and 100.

The method to calculate the similarity between a research document and a collection of documents (research objects or publications) is based on the semantic metadata and text content. The semantic metadata comes from the document enrichment service (see section 4.1), in particular the main lemmas and main concepts. Other key terms are obtained from titles and descriptions by selecting the words with the highest TF-IDF¹⁵ scores.

Figure 2 shows the main components that are used to calculate the novelty of a research object. Each number represents the information that is passed to each component, with the following meanings:

1. Request with RO id
2. RO id
3. Title, description, key lemmas, key concepts
4. Key terms (title, description), key lemmas, key concepts
5. Similar ROs
6. Key terms (title, description), key lemmas, key concepts
7. Similar publications
8. NoveltyScore, similar ROs and similar publications

¹⁵ <https://en.wikipedia.org/wiki/Tf%E2%80%93idf>

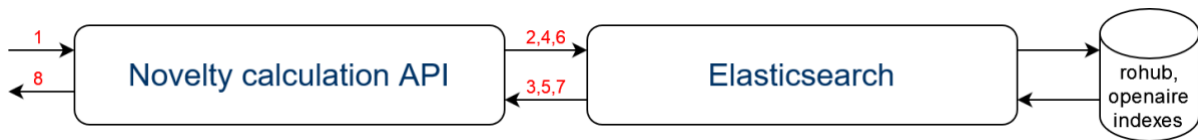


Figure 2 Novelty score service diagram

As it is shown in the diagram, we use Elasticsearch¹⁶ to store the research objects, the OpenAIRE publications and their metadata, including the title, description, key lemmas, and key concepts, that are used to calculate the novelty score. The novelty calculation API is a REST service that can be called with a POST request to the URL https://reliance.expertcustomers.ai/extended/novelty_calculation with a JSON as request body. The body must contain the research object id, as example:

POST /extended/novelty_calculation HTTP/1.1

Host: reliance.expertcustomers.ai

Content-Type: application/JSON

Content-Length: xx

```
{
  "id": "https://w3id.org/ro-id/6bc62582-2f11-4983-a51c-0af32459eca6"
}
```

If the research object exists, the response will be a JSON containing the novelty score, the similar research objects, and the similar publications, if there are any. As for example:

```
{
  "id": "https://w3id.org/ro-id/6bc62582-2f11-4983-a51c-0af32459eca6",
  "noveltyScore": 40.8846021576463,
  "similarROs": [
    {
      "id": "https://w3id.org/ro-id/94486a7f-e046-461f-bbb9-334ec7b57040",
      "smilarityScore": 0.5896301393569083,
      "title": "Tree crown delineation using detecttreeRGB (Jupyter Notebook) published in the Environmental Data Science book",
      "description": "The research object refers to the Tree crown delineation using detecttreeRGB notebook published in the Environmental Data Science book.",
      "sharedTerminology": {
        "concepts": [
          "<em>research</em>"
        ],
        "description": [
          "The research object refers to the <em>Tree</em> <em>crown</em> <em>delineation</em> using <em>detecttreeRGB</em> <em>notebook</em> <em>published</em> in the"
        ],
        "title": [
          "<em>Tree</em> <em>crown</em> <em>delineation</em> using <em>detecttreeRGB</em> (<em>Jupyter</em> <em>Notebook</em>) <em>published</em> in the <em>Environmental</em> Data <em>Science</em>",
          "<em>book</em>"
        ]
      },
      ...
    }
  ],
  ...
}
```

¹⁶ <https://www.elastic.co/elasticsearch/>

```

    ],
    "similarPublications": [{...}]
  }

```

Each similar research object or similar publication contains, in addition to the similarity score, the shared terminology with the research object that has been evaluated. Such shared terminology includes the shared concepts (main lemmas and main concepts), and the key terms from the description and from the title that are in both research works.

3.4 Reading Comprehension: Question Generation

The reading comprehension service helps users to better understand the content of a research object by challenging them to test their understanding and providing answers to the proposed questions.

The input of the service is the title and description of a research object. This text is injected into the question generation (QG) module. Then the generated questions and the text is fed into the question answering (QA) module. All the questions with an answer are finally returned to the user.

To generate questions, we use a T5 model (Raffel, et al., 2020) and a BART model (Lewis, et al., 2020) fine-tuned on question generation¹⁷. We use two models to increase the number and diversity of questions for each text. Both T5 and BART have excelled in sequence generation tasks, such as abstractive summarization and abstractive question answering. The models we reuse were fine-tuned using SQuAD1.1 (Rajpurkar, Zhang, Lopyrev, & Liang, 2016), which consists of 100,000 questions created from Wikipedia articles where answers are segments in text passages. We reuse a T5 large model and Bart base model since both models achieved the highest evaluation metrics¹⁸ when evaluated using SQuAD1.1.

T5 is fine-tuned using an answer-aware approach where the model is presented with the answer and a passage to generate the question. T5 is trained on a multitask objective to i) extract answers, ii) generate questions for answers using passages as context, and iii) extract answers for the generated questions. The answer for the generated question is compared with the answer used to generate the questions. BART is fine-tuned following an answer-agnostic approach where the model is trained to generate questions from passages without information about the answers. During generation, we use beam search as decoding method, with 5 as number of beams. Beam search keeps the most probable sequence of words at each time step and chooses the final sequence that has the overall highest probability.

To answer the questions that have been generated, we use a RoBERTa model, fine-tuned for extractive question answering in SQuAD2.0 (Rajpurkar, Jia, & Liang, Know what you don't know: Unanswerable questions for SQuAD, 2018). SQuAD2.0 adds 50,000 unanswerable questions to SQuAD1.1. Thus, the fine-tuned RoBERTa can generate answers or not depending on the question. We choose RoBERTa since it achieves a very competitive performance of 89.8 f-score (Liu, et al., 2019) when evaluated on SQuAD2.0.

The question generation service can be used by sending a POST request to the URL https://reliance.expertcustomers.ai/extended/question_generation with a JSON as content body with the research object id, as example:

POST /extended/question_generation HTTP/1.1

Host: reliance.expertcustomers.ai

Content-Type: application/JSON

Content-Length: xx

```

{
  "id": "https://w3id.org/ro-id/9bf840e3-7a39-41fc-be39-7eed9dc294db"
}

```

¹⁷ <https://paperswithcode.com/task/question-generation>

¹⁸ <https://github.com/JaquJaqu/t5-question-generation#qg-model-cards>

```
}
```

The response will be a JSON like the following:

```
{
  "id": "https://w3id.org/ro-id/9bf840e3-7a39-41fc-be39-7eed9dc294db",
  "title": "POPANE DATASET - Psychophysiology Of Positive And Negative Emotions",
  "description": "Subjective experience along with physiological activity are fundamental components of emotional responding...",
  "questions": [
    {
      "question": "What is POPANE DATASET?",
      "score": 0.6144694685935974,
      "answer": "Psychophysiology Of Positive And Negative Emotions",
      "answer_score": 0.7650960683822632
    },
    {
      "question": "What is the largest, consistent psychophysiological dataset on emotions ever collected?",
      "score": 0.7484205365180969,
      "answer": "POPANE",
      "answer_score": 0.3542005717754364
    }
  ]
}
```

For each question we provide a score that can be interpreted as the confidence of the generated question among the other generated questions in the beam search. In addition, we include the answer for the question, and its confidence score.

3.5 Enrichment Web Application

Since the enrichment web application visualizes the metadata added by the enrichment service described in section 4 and the new extended analytic services presented above, we think it is more appropriate to describe the web application at the end of section 4. Thus, we describe the enrichment web application in section 4.3.

4 Semantic annotations and information extraction: Enrichment API

The research object enrichment service adds a variety of metadata to research objects. The process starts by identifying the documents in the research object and sending them to the document enrichment service. The document enrichment service relying on different natural language processing models and tools returns metadata for the given document that then is aggregated and filtered by the research object enrichment service to produce the final set of metadata.

In the latest version of the research object enrichment, we implement an error notification mechanism with ROHub so that whenever an error is raised in the enrichment an error code and message are sent to ROHub so that the administrator of the platform can take actions accordingly. In addition, the document types supported by the document enrichment is extended to include Jupyter Notebooks and readme files in GitHub repositories. In addition, the latest version of the document enrichment service detects data cubes identifiers in the text.

4.1 Document Enrichment

The Document Enrichment processes distinct types of documents and generates metadata describing the document content and assisting in its comprehension. During the second stage of the project, as

shown in section 3, we focus on adding new features to the service that can improve the user comprehension of text content. Such features comprise the extraction of additional metadata and new processing capabilities that allow us to increase the number of document types that the system can process.

As shown in Figure 3, we cluster the metadata in the following groups:

- **Topics.** This metadata categorizes the text following a specific taxonomy, providing different perspectives regarding the text that is being analysed. In the first stage of the project, we integrated Domains that are modelled in Expert.ai Knowledge Graph that we extend in RELIANCE with vocabulary from Earth and Environmental sciences.
- **Key Elements.** The fields of this section extract the main elements of the text with different granularities. These elements try to summarize and condense in few terms the analysed text. In the first phase, we extracted concepts and phrases (called expressions back then).
- **Entities.** This section comprises all the named entities that our system is capable of extract. In the first stage we could extract people, places (now locations) and organizations.

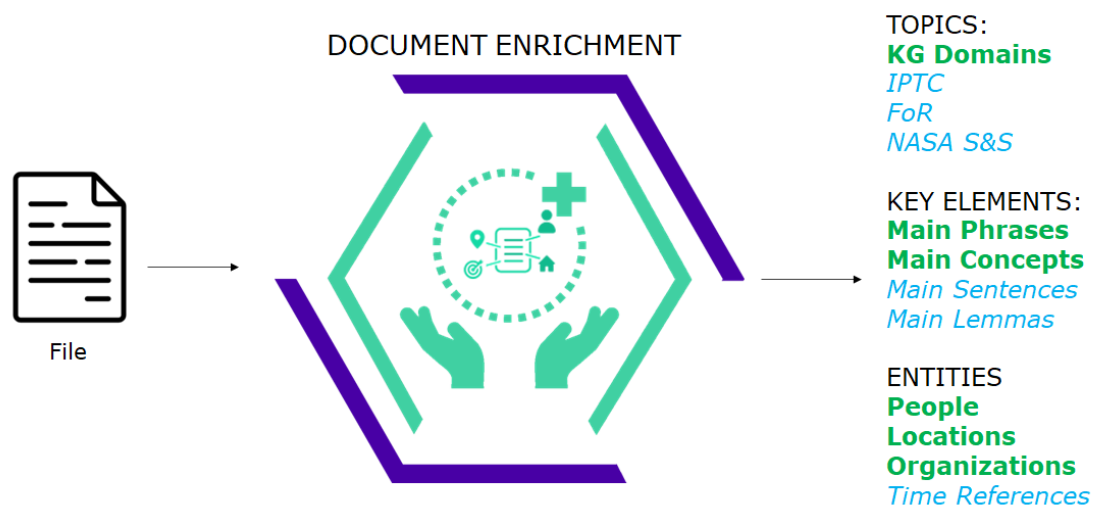


Figure 3. Summary of the Document Enrichment workflow. In green the metadata available in the first stage. In blue the metadata added in this stage.

In the second stage of the project, we added new metadata in the three metadata clusters: topics, key elements and entities. In Topics we add three new taxonomies: IPTC Media Topics¹⁹, Fields of Research (FoR)²⁰ and NASA Subjects and Scope²¹. Each one of these taxonomies (in addition to the ones already integrated during the first stage) provides a different lens to characterize the text content of a document. While IPTC is a general-purpose taxonomy, FoR and NASA Subject and Scope are centred in the scientific domain. Since in the first version of this document we describe the FoR classification model, we will focus on IPTC and the NASA Scope and subject categorization.

- The IPTC Media Topics is a three-level taxonomy based on concepts from media. It can be seen as a classification provided by a journalist or other media related analyst. This taxonomy classification is integrated in our expert.ai NLP Suite, along with the extended Knowledge Graph that we use to extract the Domain of the texts.
- The NASA Scope and Subject taxonomy counts with 11 subcategories and 88 subcategories. To develop this service, we train a transformer (Vaswani, et al., 2017) using a collection of

¹⁹ <https://iptc.org/standards/media-topics/>

²⁰ <https://www.arc.gov.au/grants/grant-application/classification-codes-rfcd-seo-and-anzsis-codes>

²¹ <https://ntrs.nasa.gov/api/citations/20000025197/downloads/20000025197.pdf>

Technical Reports from NASA²², that are categorized following the taxonomy. We train a RoBERTa large (Liu, et al., 2019) model for 4 epochs using a batch size of 2 and a learning rate of 1e05, and cross entropy as loss function. We use Expert.ai NL API²³ to get the main sentences in the text that we use to feed the model²⁴. The output of the model is the subcategory assigned to the text that can be used to also infer the main category. In Table 2 we show the evaluation results of the classifier.

Table 2. Evaluation results of the NASA Subjects and Scope classifier

NASA Subjects and Scope	Precision	Recall	F1-Score
Aeronautics	0.8044	0.7902	0.7952
Astronautics	0.7225	0.7267	0.7246
Chemistry and Materials	0.734	0.7422	0.7381
Engineering	0.739	0.7331	0.736
Geosciences	0.8636	0.8783	0.8709
Life Sciences	0.8359	0.875	0.855
Mathematical and Computer Sciences	0.7457	0.7086	0.7267
Physics	0.6909	0.6872	0.6891
Social and Information Sciences	0.6008	0.544	0.571
Space Sciences	0.8868	0.8921	0.8895
General	0.5961	0.4942	0.5404
Weighted Average	0.7908	0.7918	0.7912

In Key Elements we add a pair of new metadata of different granularity. First, the main sentences, that extracts a few sentences that can be treated as a summary of the text. Second, the main lemmas, which are the canonical form of the words or group of words with the highest importance within the text.

Finally, in entities, we add two new types of entities: time references and data cubes²⁵. Time references are dates, time ranges or anything related to a timeline (days of the week, months, seasons, etc), while our data cube extraction module locates data cube IDs within a text. We also add other references to the current entities, such as the wikidata ID (Vrandečić & Krötzsch, 2014) for people, organizations and locations, and the geonames ID²⁶ for locations, which helps to complete the information given by the entities themselves.

Regarding the new processing capabilities, our current version of the Enrichment service can process not only Word documents, Power Point files, PDF files or plain text files, but also Jupyter Notebooks and GitHub repositories. In the case of Jupyter Notebooks, we extract the text from the markdown sections of the document. Regarding the GitHub repositories, we try to access to the readme file in order to extract the more relevant text information from them.

4.2 Research Object Enrichment

The research object Enrichment acts as a resource extractor and metadata aggregator. This service extracts the resources from a research object that can be processed by the Document Enrichment, and aggregates the metadata generated using as input the collection of selected documents. During this phase we focus in three main points: 1) adapt this service to the new metadata added by the

²² <https://www.sti.nasa.gov/harvesting-data-from-ntrs/>

²³ <https://docs.expert.ai/nlapi/latest/>

²⁴ RoBERTa processes text up to 512 tokens

²⁵ <https://www.ogc.org/projects/initiatives/gdc>

²⁶ <https://www.geonames.org/>

Document Enrichment, 2) provide ROHub managers with more information relative to the status of each request 3) change the output of the service to a more developer-friendly format such as JSON. To enhance the communication between ROHub and the enrichment service, we have added two new fields to the service output: “status” and “status code”. These two fields provide information regarding the status of the request made by the user. We define a list of status codes that provides specific feedback about the execution of each one of the modules of the enrichment service: ROHub login, metadata extraction, resource selection and document enrichment. If one of the modules fails, the service returns the corresponding code. If the request has been properly processed, the service will return a code 0. The list of status codes and corresponding message is the following:

- Code 0: Correct Enrichment.
- Code 1: Error trying to login in ROHub
- Code 2: Error trying to obtain metadata
- Code 3: Error trying to select resources
- Code 4: Error during document enrichment.

In the first version of enrichment service the output was in Turtle format. We have replaced Turtle with a JSON format which provides a handier and more developer-friendly summary of the annotations generated by the service. We can see the JSON file specification below²⁷.

```
{
  "id": <string>,
  "title": <string>,
  "description": <string>,
  "creator": <string>,
  "created": <string>,
  "author": <string>,
  "sketch": <string>,
  "status": <string>,
  "status_code": <integer>,
  "topic_domains": array [{"topic": <string>, "score": <float>, "normScore": <float>}],
  "topic_iptcs": array [{"topic": <string>, "path": <string>}]
  "topic_fors": array [{"topic": <string>, "subtopic": <string>, "score": <float>, "normScore":
<float>}],
  "topic_nasa": array [{"topic": <string>, "subtopic": <string>, "score": <float>, "normScore":
<float>}],
  "ke_sentences": array [{"keyElement": <string>, "score": <float>, "normScore": <float>}],
  "ke_lemmas": array [{"keyElement": <string>, "score": <float>, "normScore": <float>}],
  "ke_phrases": array [{"keyElement": <string>, "score": <float>, "normScore": <float>}],
  "ke_concepts": array [{"keyElement": <string>, "score": <float>, "normScore": <float>}],
  "entity_organizations": [{"entity": <string>, "wikidata": <string>}],
  "entity_locations": [{"entity": <string>, "wikidata": <string>, "geoname": <string>}],
  "entity_timerefs": [{"entity": <string>}],
  "entity_people": [{"entity": <string>, "wikidata": <string>}],
  "entity_datacubes": [{"entity": <string>}]
}
```

4.3 Document Enrichment Demo

The Document Enrichment Demo is a web application that provides researchers with a complete visualization of the information extracted and generated from scientific text using the Document

²⁷ The score in the topics_for and topics_nasa indicates the confidence of the corresponding classifiers. The other scores are indicating the importance of such metadata piece in its own metadata category.

Enrichment service. It can be referred as the visual component of the Enrichment, and it helps to understand, aggregate, and analyse the enhanced metadata generated by the service. It can be accessed through the URL <https://reliance.expertcustomers.ai/enrichment>.

Document Enrichment Demo

Enhanced annotations from documents



Select one of the sample documents or submit your own text and click **ANALYZE**

Example 1: Campi Flegrei 2011-2012 dMODELS

[Document source](#)

We have developed a MATLAB software package for the most common models used to interpret deformation measurements near faults and active volcanic centers. The emphasis is on analytical models of deformation that can be compared with data from the Global Positioning System (GPS), InSAR, tiltmeters and strainmeters. Source models include pressurized spherical, ellipsoidal and sill-like magma chambers in an elastic, homogeneous, flat half-space. Dikes and faults are described following the mathematical notation for rectangular dislocations in an elastic, homogeneous, flat half-space. All the expressions have been checked for typographical errors that might have been present in the original literature, extended to include deformation and strain within the Earth's crust (as opposed to only the Earth's surface) and verified against finite element models. A set of GPS measurements from the 2006 eruption at Augustine Volcano (Alaska) is used to test the software package. The results show that the best fit source to the GPS data is a spherical intrusion, about 880 m beneath the volcano's summit.

1103/10000

Analyze

Figure 4. Screenshot of the upper part of the Document Enrichment Demo

In the upper part of the application, as shown in Figure 4, we can see a dropdown menu with a list of text examples. Each one of these examples have been selected from research objects hosted by ROHub. Under the dropdown menu we can find a hyperlink to the source of the text. Once an example is selected, the text box below is filled with the text content of the example. Moreover, the content of the text box can be modified or rewritten by the user to analyse their own pieces of text. The only limitation for the text is to have less than 10.000 characters. Once the button “Analyse” is clicked, the Document Enrichment service is called with the content of the text box, and the lower part of the screen will show the generated metadata within the different sections, using the left panel “Choose your mode”.

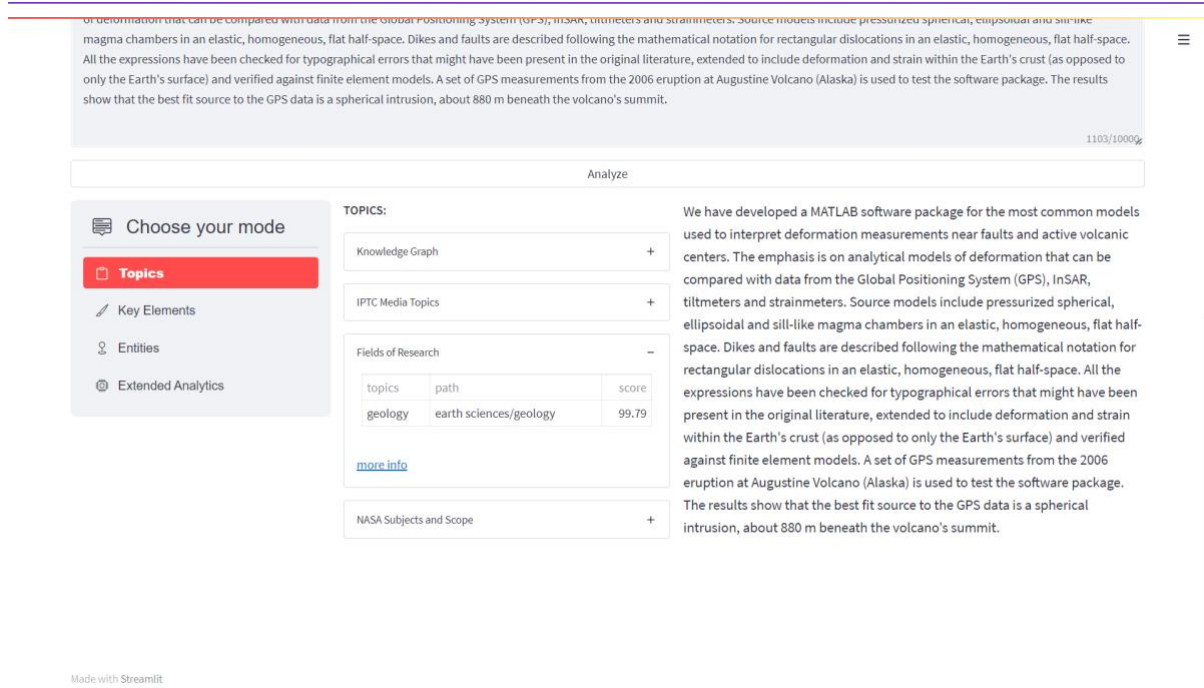


Figure 5. Screenshot of the Topics section in the Document Enrichment Demo

In the Topics section (Figure 5), you can unfold in the central panel the classification of the example text following the different taxonomies offered by the service, along with a confidence score given by the models involved.

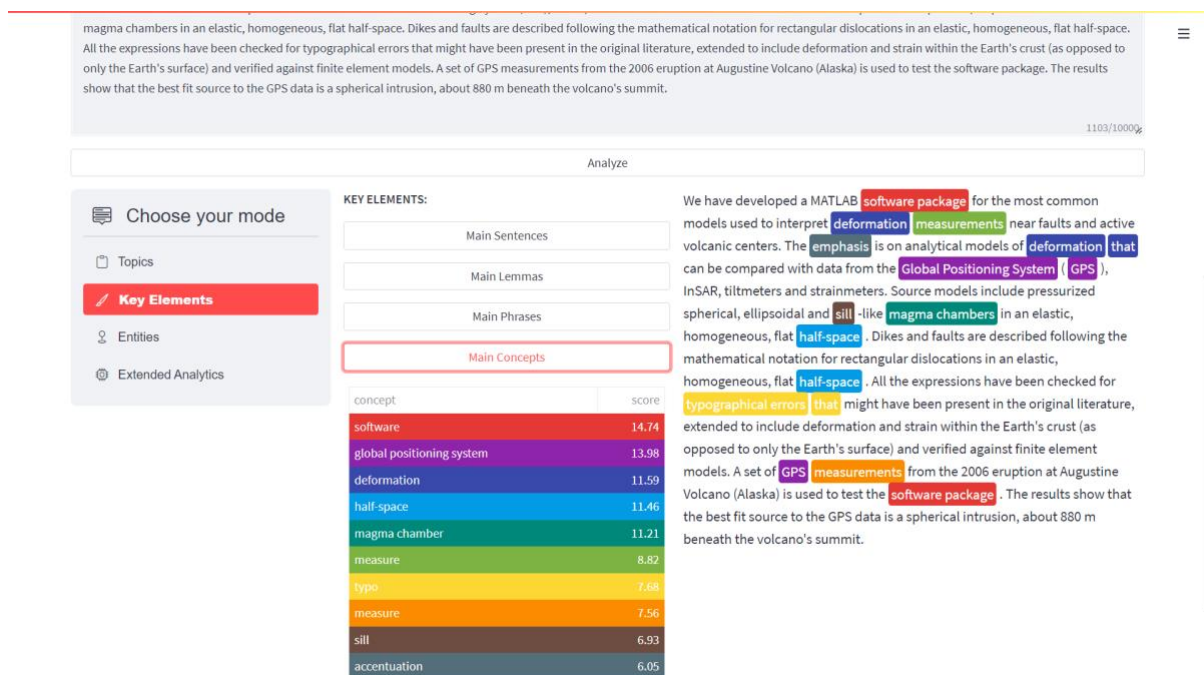


Figure 6. Screenshot of the Key Elements section in the Document Enrichment Demo

In the Key Elements section (Figure 6), each type of element can be unfolded by clicking on it in the central panel. The elements appear using a palette of colours, and the text is highlighted using the same colours to easily identify them in their textual context. There is also available a score value associated to each element which represents their importance within the analysed text as a percentage.

measurement data from almost 2000 monitoring stations during 2015-2019 and then applied to the same stations in 2020, providing predictions of expected concentrations in the absence of a lockdown. The difference between the modelled levels and the actual measurements from 2020 is used to calculate the impact of the lockdown measures adjusted for confounding effects, such as meteorology and temporal trends. The study is focused on April 2020, the month with the strongest reductions in NO₂, as well as on the gradual recovery until the end of July. Significant differences between the countries are identified, with the largest NO₂ reductions in Spain, France, Italy, Great Britain and Portugal and the smallest in eastern countries (Poland and Hungary). The model is found to perform best for urban and suburban sites. A comparison between the found relative changes in urban surface NO₂ data during the lockdown and the corresponding changes in tropospheric vertical NO₂ column density as observed by the TROPOMI instrument on Sentinel-5P revealed good agreement despite substantial differences in the observing method

1520/10000

Analyze

Choose your mode

- Topics
- Key Elements
- Entities**
- Extended Analytics

entities	type	LOD
Italy	location	
United Kingdom	location	
France	location	
Hungary	location	
Poland	location	
Portugal	location	
Spain	location	
Europe	location	
from almost 2000	time reference	
during 2015-2019	time reference	
in 2020	time reference	
from 2020	time reference	
on April 2020	time reference	

In this paper, the effect of the lockdown measures on nitrogen dioxide (NO₂) in Europe is analysed by a statistical model approach based on a generalised additive model (GAM). The GAM is designed to find relationships between various meteorological parameters and temporal metrics (day of week, season, etc.) on the one hand and the level of pollutants on the other. The model is first trained on measurement data from almost 2000 monitoring stations during 2015-2019 and then applied to the same stations in 2020, providing predictions of expected concentrations in the absence of a lockdown. The difference between the modelled levels and the actual measurements from 2020 is used to calculate the impact of the lockdown measures adjusted for confounding effects, such as meteorology and temporal trends. The study is focused on April 2020, the month with the strongest reductions in NO₂, as well as on the gradual recovery until the end of July. Significant differences between the countries are identified, with the largest NO₂ reductions in Spain, France, Italy, Great Britain and Portugal and the smallest in eastern countries (Poland and Hungary). The model is found to perform best for urban and suburban sites. A comparison between the found relative changes in urban surface NO₂ data during the lockdown and the corresponding changes in tropospheric vertical NO₂ column density as observed by the TROPOMI instrument on Sentinel-5P revealed good agreement despite substantial differences in the observing method

Figure 7. Screenshot of the Entities section in the Document Enrichment Demo

The entity section (Figure 7) shows in the central panel the list of entities found in the text. The user can identify the type of entity looking at the “type” column or checking the colour palette used for them (blue for people, red for locations, orange for time references and green for organizations). The Linked Open Data “LOD” column displays the hyperlink to the associated Wikidata and Geonames pages to each entity if they have them available. As in the Key Elements section, the right panel displays highlighted the entities with the same colour palette as the central panel.

Choose your mode

- Topics
- Key Elements
- Entities
- Extended Analytics**

EXTENDED:

Claims

verified claims	source
The coronavirus disease (COVID-19), a variant of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) originated in Wuhan city of China and has now transmitted over the world.	
The new coronavirus disease 2019 (COVID-19) severe acute respiratory syndrome coronavirus 2 was first discovered in Wuhan (China) in December 2019 and belongs to the same family as that of the severe acute respiratory syndrome coronavirus 1.	
The novel SARS-CoV-2 outbreak was declared as pandemic by the World Health Organization (WHO) on March 11, 2020.	
This cluster quickly spread across the globe and led the World Health Organization (WHO) to declare severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) a pandemic on March 11, 2020.	

Questions and Answers

Challenges and Solutions

Coronavirus disease 2019 (COVID-19), first reported in Wuhan, the capital of Hubei, China, has been associated to a novel coronavirus, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). In March 2020, the World Health Organization declared the SARS-CoV-2 infection a global pandemic. Soon after, the number of cases soared dramatically, spreading across China and worldwide. Italy has had 12,462 confirmed cases according to the Italian National Institute of Health (ISS) as of March 11, and after the 'lockdown' of the entire territory, by May 4, 209,254 cases of COVID-19 and 26,892 associated deaths have been reported. We performed a review to describe, in particular, the origin and the diffusion of COVID-19 in Italy, underlying how the geographical circulation has been heterogeneous and the importance of pathophysiology in the involvement of cardiovascular and neurological clinical manifestations.

Figure 8. Screenshot of the Extended Analytics section in the Document Enrichment Demo

The last section of the Demo shows the output of three extended analytic services for the example text: Claim Analysis, Question Generation and Challenges and Solutions. As seen in Figure 8, after clicking in “Claims”, the application will highlight the extracted claims from the text in the right panel, and with the same colour, it will display in the central panel their associated claim from the Scientific literature, along with a hyperlink to the source. The Question and Answers menu displays in the central panel the generated questions, with the score given by the Question Generation model. In the right panel the answer will be highlighted. Finally, the Challenges and Solutions menu will display solutions in a green colour and challenges in a red colour, in both central and right panels.

5 Enhanced Content-based Retrieval

In the first stage of the project, we developed and integrated in EOSC services supporting the search and recommendation of research objects. In addition, we presented the Collaboration Spheres web application that leverages the recommendation service to allow user to get recommended the research objects in an easy-to-use graphical interface.

In the second stage we implement the RELiance daSHboard that offers a content-oriented high-level overview of the research object collection in ROHub. Remarkably, the dashboard is interactive and allows refining the research object collection using advanced search capabilities in the search box and by interacting with the widgets used to depict the metadata distribution. This means that RELISH can be used as the GUI of the search engine service.

5.1 RELISH: RELiance daSHboard

RELISH stands for RELiance daSHboard, and it is a tool which allows the user to navigate and have a high-level content-oriented overview of the collection of research objects in ROHub. It is built using the Elastic stack including Elasticsearch and Kibana²⁸. The dashboard relies on the same index used for the Search and Recommendation API. This dashboard exploits the semantic metadata and provides a visualization about how the collection of research objects is characterized, allowing the user to select those parts that are relevant to them, thanks to the interaction with the tool and its filters.

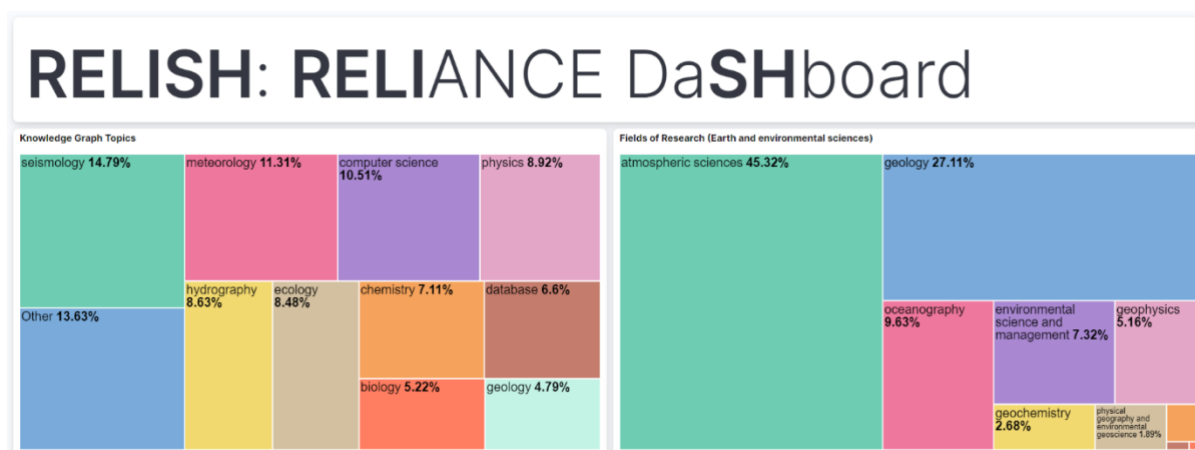


Figure 9. Screenshot of RELISH

RELISH is an interactive web application (Figure 9) that includes two main components: the action bar and the panels. The user can navigate through both components and interact with them, customizing

²⁸ <https://www.elastic.co/es/kibana/>

their view to obtain specific aspects from part or from the whole collection, such as the distribution of topics or the most used concepts.

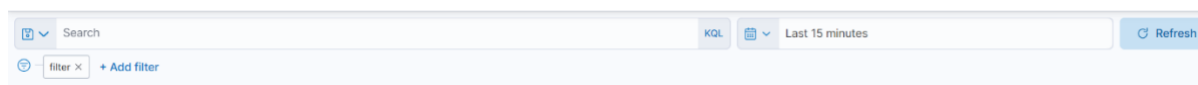


Figure 10. Screenshot of the action bar from RELISH

The action bar (Figure 10) allows the user to perform queries against the collection of enriched and basic metadata from research objects, giving them access to a series of filters, such as time ranges, authors, concepts, etc., that will automatically change the view of the panels. Thus, the query box is the component where users can perform either single term queries or faceted queries.

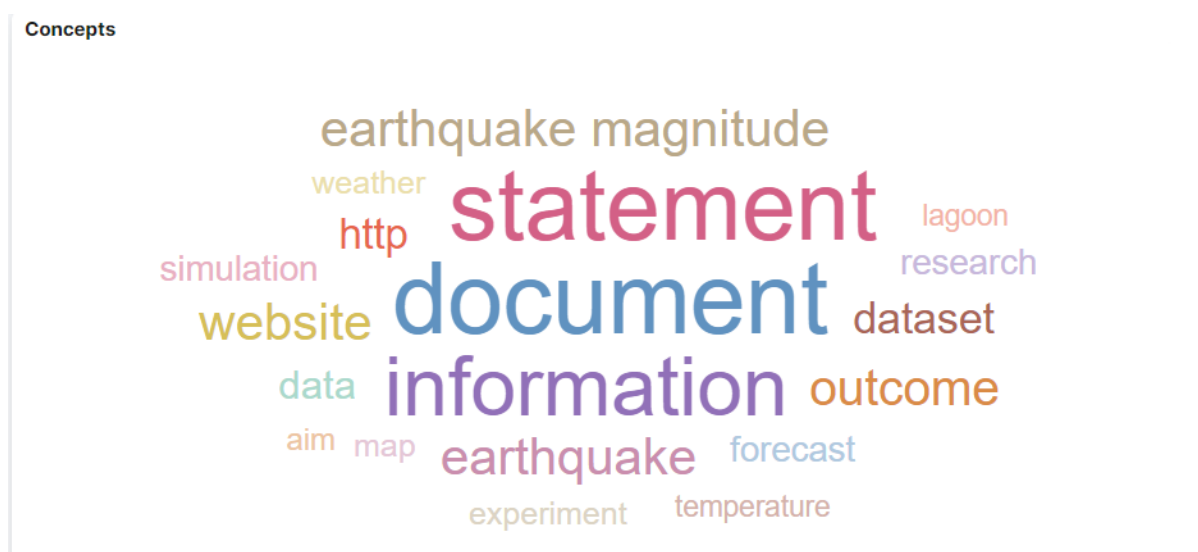


Figure 11. Screenshot of the Concepts panel from RELISH

The panels are the visual component of RELISH. There are currently two types of panels in RELISH: treemaps (Figure 9) and clouds (Figure 11). The treemap shows the distribution of the collection following a single field and the cloud of terms is a size-relative visualization of the main key elements of the collection. Both are interactable, and a click in one of the elements will filter the query to look for research objects which contain that element.



title.keyword: Ascending	Last id
IranQuakeNov2017@IGARSS	https://w3id.org/ro-id/fee98875-b261-4be5-8182-cc26e63b4e2c
(2021) Integrated Water_Vapour on the Thule Station. MODIS data _(TEST)	https://w3id.org/ro-id/36113393-cbf6-4bac-8698-d69cf6cd0329

Figure 12. Screenshot of the lowermost part of RELISH with the filtered list of research objects

Finally, in the lowermost part of the Dashboard (Figure 12), it is possible to get the collection of research objects that fulfil the query conditions and filters applied.

6 Conclusions and future work

The RELIANCE Text mining and enrichment services generate metadata useful to increase the findability and ease the comprehension of research objects and in general of text documents in the scientific domains of interest for the RELIANCE user communities. The metadata generated by the

enrichment service is comprehensive and includes concepts, phrases, entities and their types, time references, main sentences, data cube identifiers, topics and categories from diverse taxonomies such as IPTC, Nasa Subjects and Scope, Springer Nature's Fields of Research, and domains from the Expert.ai knowledge graph extended with scientific terminology. Moreover, the enrichment service supports different document formats, including txt, word, and pdf documents, as well as Jupyter Notebooks. We leverage the metadata added to research objects along with the text through a search engine and a recommender system to carry out search and content-based recommendation beyond keywords taking into account the meaning of words encoded in the concepts. The enrichment, search engine, and recommendation services are integrated in EOSC.

In addition, the services oriented to ease the comprehension of research objects and documents identify scientific claims in their text and link them to related claims made in other research work and their corresponding publications, extract challenges and solutions, generate a list of questions and their associated answers about the research description, and assign a novelty score by analysing related publications and research work. These services apply advanced natural language processing capabilities such as text generation, machine reading comprehension, semantic similarity, and text classification that are still open research problems.

We have also made available the collaboration spheres web application, which builds upon the recommendation service, and visualization tools such as the enrichment web application, where users can inspect the results produced by the enrichment service. In addition, the Research Object dashboard RELISH provides a visualization of the research object collection in ROHub based on the metadata extracted by the text mining services, as well as a search and exploration mechanism.

As future work we foresee natural language processing tools going beyond the current analysis of text to assist scientists in their research endeavours. For example, a natural step is to support or refute scientific claims and offer bibliographic references to justify the decision. In addition, language models pretrained for text generation in the scientific domain with access to scholarly communications can help to describe the state of the art on a particular scientific field and task and propose new research questions learning from existing literature.

7 References

- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . . Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., . . . Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*.
- Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.
- Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv*. doi:10.48550/ARXIV.1606.05250
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems (NIPS)*.
- Wright, D., & Augenstein, I. (2021). CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding. *ACL-IJCNLP*.
- Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *SIGIR: Special Interest Group on Information Retrieval*.

- Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. *International Workshop on Semantic Evaluation (SemEval)*.
- Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10), 8-85.