Scientific Lake

DOI: 10.5281/zenodo.13239445

Deliverable 4.1: Initial version of the smart reproducibility assistance service

Due Date of Deliverable	30/06/2024
Actual Submission Date	30/06/2024
Work Package	WP4
Tasks	T4.1, T4.2, T4.3, T4.4
Type	OTHER
Approval Status	Accepted
Version	v1.0
Number of Pages	41

Abstract

This deliverable report outlines the progress made in Work Package 4 (WP4) of the SciLake project, focusing on the development of the initial version of the Smart Reproducibility Assistance Service. This service aims to contribute in tackling the reproducibility crisis in scientific research by leveraging the rich data within the Scientific Lake. By enhancing Scientific Knowledge Graphs (SKGs) with critical information on research outputs, the service is aimed at helping researchers to identify reproducible studies and understand the extent to which findings have been replicated.

This project has received funding from the European Union's Horizon Europe framework programme under grant agreement No. 101058573. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the European Research Executive Agency can be held responsible for them.

Revision history

Version	Date	Reason	Revised by
0.0	9/5/2024	Agreement on structure	César Parra Rojas
0.1	8/6/2024	First Draft	César Parra Rojas
0.2	10/6/2024	Intermediate version	César Parra Rojas
0.3	10/6/2024	Ready for Peer review	Pablo Accuosto
0.4	27/6/2024	Peer review comments addressed	Pablo Accuosto
1.0	29/6/2024	Final Version after proofreading	César Parra Rojas

Author List

Organisation	Name	Contact Information
SIRIS	César Parra Rojas	cesar.parra@sirisacademic.com
SIRIS	Pablo Accuosto	pablo.accuosto@sirisacademic.com

Contributor List

Organisation	Name	Contact Information
ARC	Sotiris Kotitsas	sotiris.kotitsas@athenarc.gr
ARC	Haris Papageorgiou	haris@athenarc.gr
ARC	Thanasis Vergoulis	vergoulis@athenarc.gr
DFKI	Julian Moreno Schneider	julian.moreno_schneider@dfki.de
TUE	Nikolay Yakovets	n.yakovets@tue.nl
ICM	Marek Horst	mhorst@icm.edu.pl
ARC	Sokratis Sofianopoulos	s_sofian@athenarc.gr

Table of Contents

List of Tables

List of Figures

Abbreviation List

DOI	Digital Object Identifier
ROR	Research Organization Registry
NER	Named-entity Recognition
NEN	Named-entity Normalisation
GDD	Graph Generating Dependency
HDFS	Hadoop Distributed File System
KG	Knowledge Graph
LLM	Large Language Model
LoRA	Low Rank Adaptation
NMT	Neural Machine Translation
PMID	PubMed Identifier
RA	Research Artefact
RAA	Research Artefact Analysis
SKG	Scientific Knowledge Graph
TSV (file)	Tab Separated Values (file)
UI	User Interface

1. Executive Summary

In this report, we describe the initial version of the Smart Reproducibility Assistance Service, developed as part of the SciLake project. This service leverages the contents of SciLake's SKGs and the various technologies provided by SciLake's Scientific Lake architecture. Its main aim is to address the reproducibility crisis in scientific research by assessing and enhancing the reproducibility and replicability of research outputs. Transparency of methods in research ensures that all procedures, materials, and analytical techniques are clearly described,

allowing others to follow the same steps and achieve similar results, reducing the risk of errors and biases and allowing other researchers to reanalyze data, verify results, and potentially uncover new insights or corrections. This is hindered by the lack of a standardised manner in which resources other than scientific publications are cited in other works, being most of the time "hidden" in footnotes or the body of a text—as opposed to appearing as structured references—making it difficult for others to replicate studies accurately when navigating vast amounts of research outputs. In addition to this, the quality of scholarly research cannot be reliably assessed based on citation numbers alone, since they do not account for the context or content of the citations. High citation counts can result from factors unrelated to research quality, such as the popularity of a topic, self-citations, or citations in negative contexts; and some influential but niche research may receive fewer citations despite its high quality and impact within a specialised field.

In light of the above, the Smart Reproducibility Assistance Service provides functionalities to help researchers identify reliable studies by evaluating their reproducibility and replicability, by employing advanced text mining techniques to uncover missing links between different types of research outputs in the SKGs in addition to analysing the context in which citations between outputs occur—all in a multilingual framework—in order to provide an overview of the resources created and/or (re)used by a given work, and of the reception of its findings among the scientific community. This has the goal of allowing researchers to (a) verify the reliability of scientific findings, (b) discover robust and repeatable research outputs, and (c) improve transparency and rigour in scientific publishing. The service architecture is designed for it to be adaptable to different scientific domains. Within the project, service functionalities are tailored to the specific domains of the project pilots in order to assess them in real-life scenarios.

In the following sections, we provide a comprehensive overview of the Smart Reproducibility Assistance Service. Section 2 discusses the reproducibility crisis, providing background, motivation, and the challenges addressed by this service, while Sections 3 and 4 detail the design and implementation of the major components, respectively. These include research object recognition, named-entity recognition and normalisation, link recommendation, article segmentation and translation, citation-context assisted replication assessment, and the integration of these components into reproducibility badges. Finally, Section 5 summarises the report and discusses the next steps in the ongoing development and enhancement of the service.

2. Introduction

Current scientific research moves at a vigorous pace, producing vast amounts of rich, valuable knowledge from a plethora of sources; however, this rapid proliferation of information has some potential drawbacks. The number of manuscripts submitted for peer review is growing more and more every year, and the resulting increased workload for editors and reviewers, who are tasked with the in-depth scrutiny of the soundness and clarity of the methods and reported findings, has the potential to affect the overall rigour of the process. As a result, low-quality research may end up reaching the publication stage.

The traditional peer review system itself is not without its flaws, being plagued by issues of bias amplification and lack of transparency, to name a few, and a number of alternatives have been proposed, including continuous post-publication review. This, and the proliferation of preprint repositories, results in researchers facing not only an enormous scale of potentially valuable information, but also, in many cases, without the "seal of approval" of formal peer review. Furthermore, the production of scientific knowledge can take many forms, and publications are only one item in an ever-expanding list of research outputs that researchers can build upon.

The sheer scale of information makes it difficult for researchers to navigate the scientific landscape and identify high-quality, credible works. This is especially relevant in the context of the "reproducibility crisis" affecting research [1], with a large number of studies that are not amenable to verification due to lack of transparency or issues of availability—details buried behind paywalls, data not released—while a few others report conclusions of dubious validity. One of the aspects contributing to the former is the lack of a standardised manner in which to cite research outputs other than publications—e.g., datasets, software—so that the connections between these heterogeneous sources are made apparent in vast collections of data. Recent attempts at addressing this issue include repositories that endow different types of non-paper research outputs with a persistent identifier, thus making them citable in a traditional fashion—e.g., Zenodo. However, despite these efforts, most of these references remain hidden in footnotes or in direct textual mentions. This makes identifying the connection between scientific findings and a given resource a very difficult task, and hinders a thorough understanding of resource use, reuse, and impact.

Here, we present the first version of the Smart Reproducibility Assistance Service of the SciLake project. This service aims to address the technical challenges in the assessment of research reproducibility. It leverages the contents of the Scientific Lake to enrich the SKGs with valuable information in the form of new connections between research outputs. This facilitates knowledge discovery by highlighting research outputs that are more likely to be reproducible and shows the extent to which their findings have been replicated. This

information is key for researchers to get a quick overview of the scientific works they may use as a base for their own studies. This is done through the application of advanced text mining and link prediction techniques, in addition to citation context analysis.

It is important to mention that some of the presented technologies, such as research object recognition, link recommendation, article segmentation and named-entity recognition can be integrated into the broader scientific lake infrastructure. By doing so, these technologies can provide similar functionalities across different domains of scholarly content, enhancing the overall capability of the scientific lake to support the needs of diverse research communities.

3. Design

In this section we briefly discuss the main requirements of the Smart Reproducibility Assistance Service. More details about the design process can be found in Deliverable D1.2 ("Initial integrated system").

3.1. Main requirements

The aim of this service is to address the reproducibility crisis in scientific research by assessing and enhancing the reproducibility and replicability of research outputs. The service aims to provide researchers with tools to better understand and pursue reproducibility in scientific findings by leveraging the contents of SKGs within the Scientific Lake. More specifically, the service should be able to support the following main functionalities:

- Research object recognition in textual data. Making available, as part of the SK services, components designed to extract mentions of datasets, software, and other domain-specific entities from scientific texts is expected to significantly contribute to enhancing research reproducibility. Through automated identification and cataloguing within SKGs, researchers can gain a detailed overview of the resources utilised in their studies. This service is aimed at facilitating rapid access to precise information about datasets and software, essential for experiment replication and result validation. Moreover, it aims at transparent documentation of critical tools and data, thereby reducing ambiguity that often impedes reproducibility efforts. These advancements can support not only the verification of individual studies but also enable meta-analyses and the synthesis of broader trends within scientific disciplines, thereby strengthening the reliability and transparency of scientific research.
- Recognition of domain-specific research objects. This functionality is required to enhance the scientific exploration capabilities of the Scientific Lake for end-users

across four project pilots by identifying domain-specific terms within research outputs stored in SKGs. By pinpointing mentions of concepts and entities relevant to the project pilots, leveraging well-established biomedical vocabularies and other NLP resources, we aim to suggest connections that enrich the respective graphs. These suggestions will undergo review by a data curator for potential incorporation. Future efforts will extend this capability to include concepts relevant to energy and transportation, with plans to develop a flexible pipeline adaptable to diverse research communities over time.

- Research object link recommendation. The identification of missing links between research objects and research stakeholders plays a crucial role in enabling scientific reproducibility. By filling gaps in existing SKGs, these components should ensure that all pertinent resources and contributors are thoroughly documented and interconnected, so it is possible to trace the lineage and context of each research output, providing a detailed account of data collection, analysis, and interpretation processes. This is expected, in turn, to support compliance with Open Science standards and to strengthen the credibility and impact of research supported by funding bodies and institutions.
- Article segmentation for multilingual articles. The article segmentation component aims to automatically analyse the structure of scientific documents by identifying their constituent parts. By recognising and normalising various names for sections in a scientific paper (e.g., 'Conclusions' may be equivalent to 'Conclusions and Future Work', or 'Discussion'), this component would enhance understanding documents structure and organisation, aiding researchers in navigating and extracting information from complex scientific texts efficiently.
- Citation-context assisted replication assessment. This component is required to provide detailed insights into scientific citations by analysing and classifying citances, and addressing critical questions such as the intent behind citations, the authors' stance towards cited works, and specific aspects referenced from those works. This nuanced analysis is relevant because it enables researchers to accurately replicate and validate scientific findings by understanding the precise context in which prior research is cited and utilised. By clarifying the purposes and interpretations of citations, this component can enhance transparency and reliability in research, ensuring that methodologies can be faithfully reproduced and findings independently verified.
- Reproducibility and replicability badges. The implementation of reproducibility and replicability badges within the UI of the Smart Impact-driven Discovery Service (as described in D3.1) aims to enhance research integrity and reliability. These badges serve as visual indicators that enable users to easily identify research outputs with comprehensive methodological details and verified findings, or conversely, to

recognise those with potential limitations or flaws. By providing this functionality, the badges can contribute to promote transparency and trustworthiness in scientific research. They empower researchers by allowing them to assess the status of their own work against established standards, facilitating improvements where necessary.

4. Implementation

In this section, we detail the individual components of the Smart Reproducibility Assistance Service—providing links to code repositories when appropriate—and their integration into the Scientific Lake, as well as their interplay resulting in the implementation of reproducibility measures.

4.1. Research object recognition in textual data

In this section, we elaborate on two SciLake components designed to automatically recognise and extract mentions of research objects in scientific publication manuscripts or other scientific texts. Each of the components focuses on different types of research objects. In the next sections, first, we provide background on the problem of research object recognition in textual data and explain how it can help in enhancing research reproducibility. Then, we provide the technical details related to each of the aforementioned components.

4.1.1. Background

Components that are able to extract mentions of datasets, software, and other domain-specific entities from scientific publication manuscripts or other scientific textual data can contribute in enhancing the reproducibility of research. By automatically identifying and cataloguing these entities in Scientific Knowledge Graphs (SKGs), these components provide a comprehensive map of the resources used in research works. Researchers can then quickly access detailed information about the datasets and software employed, which is essential for replicating experiments and validating results. They can also quickly identify how easy it is for other researchers to reproduce their own work. This transparency ensures that all the tools and data necessary for reproduction are clearly documented, reducing the ambiguity that often hampers reproducibility efforts. A series of reproducibility-related badges and/or indicators can be calculated using this information, offering convenience to researchers and other stakeholders.

Moreover, the respective enrichments to the SKGs can significantly streamline the research process. For instance, by identifying and linking mentions of clinical trials across different studies, researchers can gain insights into how various trials are related, compare outcomes, and identify potential inconsistencies or areas for further investigation. The same applies to datasets and software, where connections between studies using similar resources can be made. This interconnected approach not only aids in the verification of individual studies but also facilitates meta-analyses and the synthesis of broader trends within a field. Ultimately, this enhances the overall reliability and robustness of scientific research, fostering a more transparent and reproducible scientific ecosystem.

4.1.2. SciNoBo RAA tool: Recognition of Datasets & Software

The **SciNoBo Research Artefact Analysis (RAA) tool**[1] performs RAA on scientific texts to identify mentions of research artefacts (RAs) such as datasets and software. It extracts these mentions along with their associated metadata and then deduplicates them to determine the unique RAs that were referenced, reused, or created in the text (see Figures 4.1.2.1 and 4.1.2.2). The tool leverages fine-tuned large language models (LLMs) and relies on a predefined list of discipline-specific keywords, key phrases, and gazetteers[2] to detect candidate mentions of RAs.

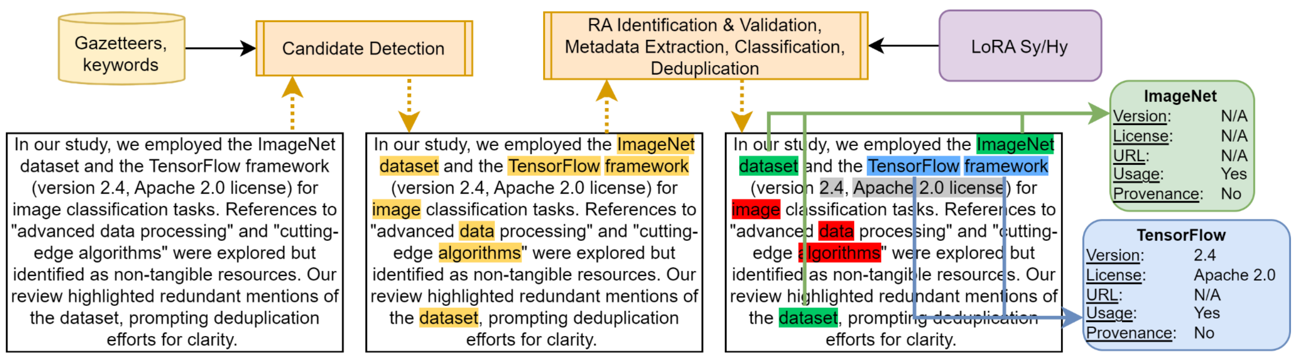

Figure 4.1.2.1: *An overview of how the SciNoBo RAA tool works.*

[1] github.com/iNoBo/scinobo-raa

[2] Lists of known entities from an external source.

Snippet	In their study, the authors utilized the PyTorch <m>library</m> (version 1.9.0) for deep learning experiments. PyTorch is released under the BSD-3-Clause license. For more information, visit https://pytorch.org/.
Type	Software
Valid	Yes
Name	PyTorch
Version	1.9.0
License	BSD-3-Clause
URL	https://pytorch.org/
Provenance	No
Usage	Yes

Figure 4.1.2.2: *An example of an RA mention containing all metadata.*

System Overview

The RAA system processes the structured text of a publication (including sentences, paragraphs, and sections with titles) to extract research artefacts (RAs). It employs a fine-tuned LLM and an optional Coreference Resolution Longformer model (SciCo)[3] [2A] for deduplication within a structured pipeline. The pipeline comprises three stages:

1. **Research Artefact Validation:**

This stage filters sentences containing research artefacts, leveraging the LLM through two subsystems:

- **Paragraph Relevance Checker**: Operates the fine-tuned LLM in a 'fast mode', evaluating whether a paragraph includes any RAs. If artefacts are detected, it marks the paragraph as worthy of further examination.

- **Research Artefact Validator**: This subsystem examines each candidate sentence for keywords utilising a candidate detection process and then classifies the candidate RA mentions as valid or merely generic references.

In the candidate detection phase, the system identifies keywords and keyphrases as triggers for datasets and software in scientific texts from a human-curated list. Additionally, gazetteers from the PapersWithCode (PwC) dataset[4] help identify

[3] SciCo Longformer model is designed for handling long scientific documents and it is used to perform hierarchical cross-document coreference resolution.
[4] github.com/paperswithcode/paperswithcode-data

candidate RA mentions that are not mentioned via these keywords, but rather the names of the corresponding datasets and software.

Regular expressions scan the text for matches of keywords, keyphrases, and gazetteers, recording the location of each mention. The system also supports a mechanism to incorporate gazetteers from external sources, enhancing the model's ability to identify named RAs in new texts and disciplines.

Then, the system utilises the fine-tuned LLM to assess the validity of each candidate RA mention by assigning a validity score against a set threshold. Only those mentions surpassing the threshold are considered valid, while the rest are considered generic references.

This validation stage ensures that only pertinent sentences are passed on for further analysis, increasing the accuracy and relevance of the extracted RAs. Figure 4.1.2.3 summarises the Research Artefact Validation pipeline.

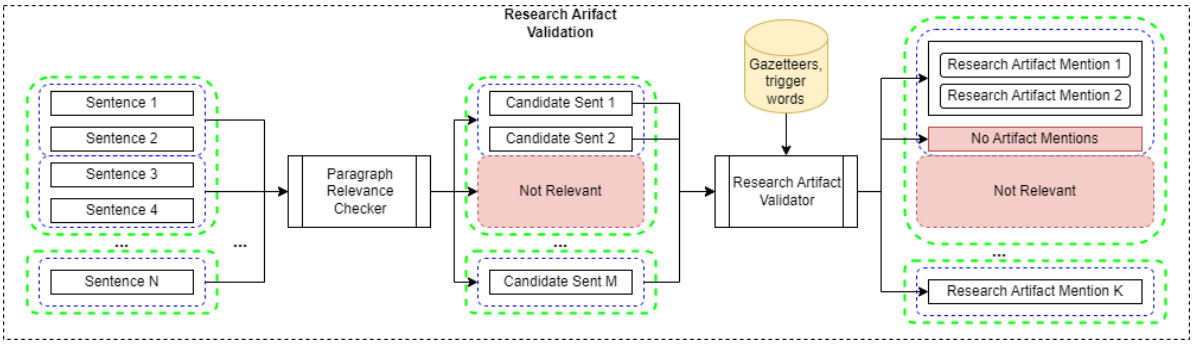

Figure 4.1.2.3: *Research Artefact Validation pipeline.*

2. Research Artefact Metadata Extraction & Classification:

This stage focuses on extracting metadata from research artefact mentions and classifying the artefacts based on ownership and usage. It leverages an LLM through two subsystems:

- **Research Artefact Metadata Extraction**: Uses the fine-tuned LLM on valid RA mentions to extract metadata such as name, version, licence, and URL from the text snippet of each valid mention.

- **Research Artefact Ownership/Usage Classification**: Uses the fine-tuned LLM to predict ownership and usage classes, determining whether the RA has been created by the authors or (re)used in their work, using a set threshold for classification.

This stage ensures comprehensive metadata capture and accurate classification of the artefacts, providing detailed insights into the usage and provenance of the RAs mentioned in the publication. Figure 4.1.2.4 illustrates the Research Artefact Metadata extraction and classification pipeline.

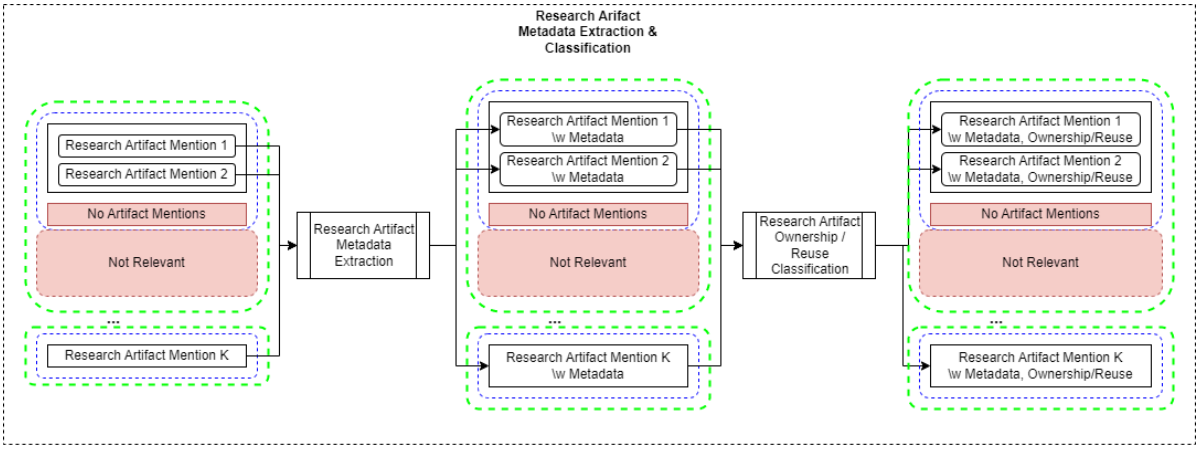

Figure 4.1.2.4: *Research Artefact Metadata extraction & classification pipeline.*

3. **<u>Research Artefact Deduplication & Re-evaluation:</u>**

This stage clusters research artefact mentions, using their context and metadata to deduplicate and reevaluate them. The process refines the metadata and classifications of "owned" and "reused" based on the consolidated context of each research artefact. It leverages the SciCo Longformer model, if available.

- **Research Artefact Deduplication & Re-evaluation**: Utilises the paragraph context and metadata of each research artefact mention to cluster them into unique research artefacts. It consolidates RA mentions by grouping them based on names and other metadata (e.g., citation marks), merging similar clusters. The optional SciCo Longformer model aids in this process by providing similarity scores for hierarchical cross-document coreference resolution. If the SciCo model is not used, deduplication is still performed based on the names and metadata of the RA mentions.

Finally, the system re-evaluates the usage and ownership classifications of the deduplicated RAs by aggregating metadata and considering the context within which each RA is mentioned, ensuring unique identification and accurate classification.

This stage ensures that each research artefact is uniquely identified and accurately classified, enhancing the reliability and clarity of the extracted information. Figure 4.1.2.5 provides an overview of the Research Artefact deduplication and re-evaluation pipeline.

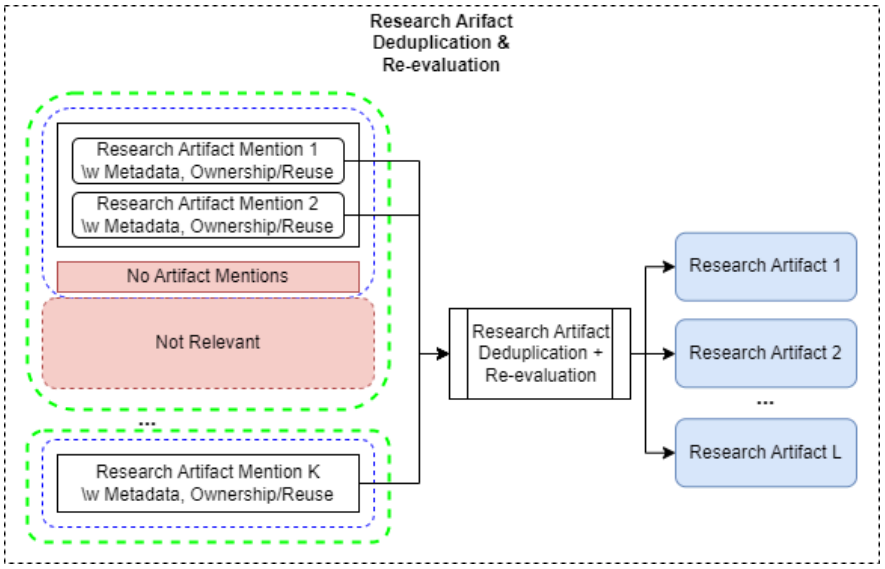

Figure 4.1.2.5: *Research Artefact deduplication & re-evaluation pipeline.*

In the current state, the tool leverages a fine-tuned version of the Flan T5 Base model[5] using the Low-Rank Adaptation (LoRA) method, called "LoRA Hy". The keyword list is tailored for the Computer Science discipline, and the gazetteer list is sourced from PapersWithCode.

Next Steps

The next steps for enhancing the Research Object Recognition in textual data component and the SciNoBo Research Artefact Analysis (RAA) tool include:

1. Fine-tuning state-of-the-art LLMs, such as Llama 3[6], for improved performance on the same task.
2. Evaluating the tool within the SciLake pilots to assess its effectiveness in real-world scenarios, assessing an indicative sample of publications.
3. Collecting and incorporating feedback from the pilots to evaluate and potentially adapt the system for other disciplines.
4. Expanding the keyword and gazetteer lists to cover a broader range of scientific fields, ensuring wider applicability and increased precision.
5. Use the mentions identified and the related metadata to enrich SciLake's SKGs.

[5] huggingface.co/google/flan-t5-base
[6] github.com/meta-llama/llama3

USEFUL RESOURCES

The next table summarises the most important resources (publications, code repositories, etc) related to the SciNoBo RAA tool.

Paper(s): • Empowering Knowledge Discovery from Scientific Literature: A novel approach to Research artefact Analysis • Streamlining Knowledge Discovery in Scientific Literature: A Comprehensive End-to-End System for Research Artefact Analysis (submitted to EMNLP 2024) **Code repo(s):** github.com/iNoBo/scinobo-raa **Docker hub:** hub.docker.com/repository/docker/intelligencenoborders/scinobo-raa/general **HF Space (Demo):** huggingface.co/spaces/iNoBo/scinobo-research-artefact-analysis

Table 4.1.2.1: *Links to code repositories and documentation of the field extraction component.*

4.1.3. Recognition of domain-specific research objects

With the objective of enriching the exploration of the scientific content of the Scientific Lake for the end-users of the four project pilots with domain-specific terms, we aim to identify mentions to concepts/entities of interest for the respective research communities within the text of the research outputs in the SKGs. The mentions identified in this way can be used to suggest potential connections for the corresponding graph, which can be subsequently reviewed by a data curator in order to decide whether to incorporate them into it.

We have started with the identification of concepts/entities relevant for the cancer and neuroscience pilots, since the biomedical domain has a plethora of well-established and standardised vocabularies/taxonomies, and NLP resources and annotated datasets abound. These capabilities will be extended to consider concepts/entities relevant for the energy and transportation pilots; in the long term, we expect to have a flexible pipeline that can accommodate other research communities.

As a first step, we follow the all-in-one named-entity recognition (AIONER) scheme [3], due to its simplicity and potential for adaptability and reusability, in order to obtain an initial version of the biomedical NER (bio-NER) module, capable of identifying a number of core biomedical entities with gold-standard datasets: genes, diseases, chemicals, species, variants, and cell lines. We retrain the original implementation based on the BioRED dataset, along with additional BioRED-consistent datasets:

- Gene: GNormPlus, NLM-Gene, DrugProt
- Disease: BC5CDR, NCBI Disease
- Chemical: BC5CDR, NLM-Chem, DrugProt
- Species: Species-800, Linnaeus
- Variant: tmVar
- Cell line: BioID

using four pre-trained language models as a base:

1. BiomedBERT-base pre-trained on abstracts from PubMed[7]
2. BiomedBERT-base pre-trained on both abstracts from PubMed and full-texts articles from PubMedCentral
3. BioLinkBERT-base
4. BioLinkBERT large

[7] Previously known as PubMedBERT, this is the original best-performing model reported on the AIONER paper.

The F1 scores of the current implementation on the BioRED test set are shown in Table 4.1.3.1.

	BiomedBERT-base abstract	BiomedBERT-base abstract+fulltext	BioLinkBERT-base	BioLinkBERT-large
Cell line	88.66	92.93	91.67	93.75
Chemical	92.61	93.04	92.97	94.01
Disease	88.89	88.40	88.17	88.90
Gene	94.82	94.59	95.84	96.02
Species	96.34	96.72	97.97	96.70
Variant	93.81	94.17	93.72	95.00
Overall	**92.28**	**92.68**	**93.14**	**93.69**

Table 4.1.3.1: *F1 scores on the BioRED test set for the four base models for biomedical entity identification trained using the AIONER scheme.*

Additional biomedical entities

In addition to the core entities above, we explore additional annotated datasets for the identification of mentions to other entities that may be of interest to the use-cases of the project. The first of these is the AnatEM corpus, containing ~1K documents manually annotated for anatomical entity mentions; in our context, we have a particular interest in the identification of, e.g., *cellular components* and *tissues*, along with *cancer* (newly-introduced in the latest release of the corpus). We extend the four aforementioned models trained using the AIONER scheme by fine-tuning them on this dataset, constrained to the selected entities. The F1 scores obtained on its test set are shown in Table 4.1.3.2.

All the data used for training and testing, the code and the four trained models are available at github.com/sirisacademic/AIObioEnts. An illustration of the output is shown in Fig. 4.1.3.1.

	BiomedBERT-base abstract	BiomedBERT-base abstract+fulltext	BioLink-base	BioLink-large
Cell component	85.00	82.54	76.72	84.25
Tissue	70.92	70.82	71.95	72.19
Cancer	87.36	84.13	88.29	86.56
Organ	74.74	76.47	77.01	81.94
Multi-tissue structure	67.87	67.36	72.77	77.96
Overall	**79.29**	**77.86**	**80.60**	**81.30**

Table 4.1.3.2: *F1 scores for the fine-tuned NER models on selected entities of the AnaTEM test set.*

Next steps

Additional data

We will continue developing the bio-NER module in order to include additional data for the current entities and/or additional entities. Currently, we have two lines of exploration:

1. Fine-tuning with the MACCROBAT dataset, containing manually-annotated clinical reports. Of particular interest among the annotated entities are *symptoms*.
2. Selecting potentially relevant entities from the controlled terms of openMINDS.

In collaboration with the representatives of the other two pilots, we will also explore the concepts of interest for their respective communities, including both existing resources and the possibility of jointly introducing datasets annotated from scratch.

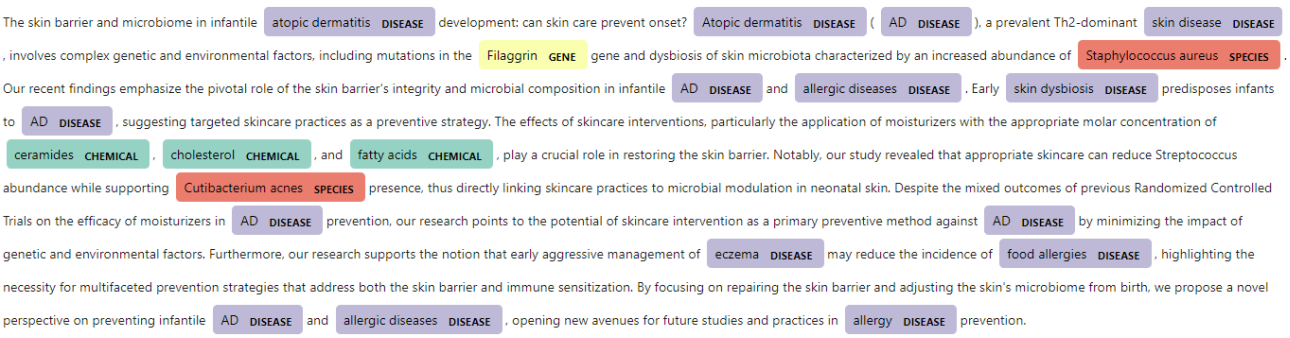

Figure 4.1.3.1: *Visualisation of the outputs of the bio-NER module on a publication title+abstract.*

For biomedical entities, the normalisation, or linking, step will be carried out using the standard biomedical taxonomies/thesauri, such as the MeSH terms. An illustration of the combined output of both the bio-NER and the (yet to be implemented) bio-NEN modules is shown in Fig. 4.1.3.2.

Other alternatives are to be considered for non-biomedical domains. In a more general sense, a domain-agnostic implementation may use other types of sources, such as WikiData[8], for the normalisation of relevant entities. Existing entity-linking models like mGENRE (multilingual GENRE[9]) and BLINK[10] can be leveraged to link identified entities to WikiData entries. Additionally, pipelines with a retrieval step based on search engines such as Elasticsearch[11], followed by a re-ranking step using either a pairwise re-ranking model and/or an LLM, will also be implemented and compared to existing solutions. SIRIS has explored entity-linking pipelines combining Elasticsearch with LLMs for an entity-linking task in another domain—disambiguating institutions mentioned in affiliation strings by means of their ROR identifiers—which has proven a valuable alternative to existing solutions.

In addition to these domain-agnostic solutions, domain-specific vocabularies can be used for the energy and transportation use cases, such as the International Nuclear Information System (INIS) thesaurus for energy and the Transport Research Thesaurus (TRT) for transportation. Solutions involving the adaptation of existing models to these domains can be explored by fine-tuning them on domain-specific corpora annotated with entities from these vocabularies.

[8] wikidata.org
[9] github.com/facebookresearch/GENRE
[10] github.com/facebookresearch/BLINK
[11] github.com/elastic/elasticsearch

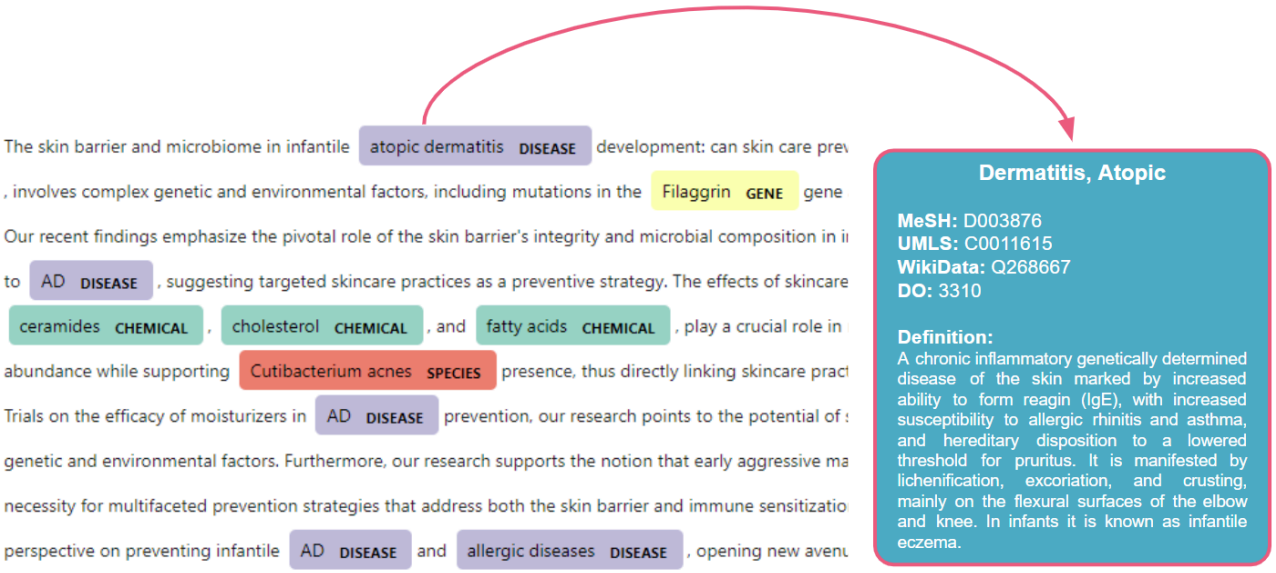

Figure 4.1.3.2: *Illustration of the expected outputs from the bio-NER and nio-NEN modules combined, on the same text as in Fig. 4.1.3.1.*

Code repo(s): github.com/sirisacademic/AIObioEnts **Model files:** huggingface.co/datasets/SIRIS-Lab/AIObioEnts-model_files

Table 4.1.3.3: *Links to code repositories and model files for the bio-NER module.*

4.2. Research object link recommendation

In this section, we introduce two SciLake components (sHINER and SciNeM) designed to recommend links between Knowledge Graph (KG) nodes, particularly research objects. These recommended links may be missing from the respective KGs and can be used to enrich them and further improve their coverage. In the next sections, first, we provide background on the problem of research object link recommendation and explain how it can help in enhancing research reproducibility. Then, we provide the technical details related to each of the aforementioned components.

4.2.1. Background

Components that are able to identify links between research objects (such as datasets, publications, and software) and research entities (such as authors, institutions, and funding sources), which are missing from existing SKGs, are invaluable for enhancing the reproducibility of scientific research. By pinpointing these missing connections, the component ensures that all relevant resources and contributors are comprehensively documented and interconnected. This comprehensive mapping makes it possible for researchers to trace the provenance and context of each research output, providing a clear and detailed account of how data was collected, analysed, and interpreted. As a result, other researchers can more accurately replicate the study's methodology and validate its findings, which is a cornerstone of scientific reproducibility.

Moreover, identifying missing links can reveal previously overlooked relationships and dependencies that are critical for understanding the full scope of a research project. For example, linking datasets to their originating experiments or software tools to their specific applications can highlight the intricate workflows involved in scientific research. This transparency not only aids in reproducibility but also fosters collaboration and knowledge sharing. Researchers can build upon existing work with greater confidence, knowing that they have a complete picture of the tools, data, and methodologies involved. Additionally, funding bodies and institutions can use these insights to ensure compliance with open science mandates and to promote practices that enhance the credibility and impact of their supported research.

Ultimately, components that provide missing links between research objects and entities, can act in a complementary fashion with those that reveal similar links by identifying mentions in scientific texts (like those presented in the previous section). These two categories of tools

combined have the potential to significantly increase the coverage of SKGs regarding the respective types of links.

4.2.2. sHINER: the Graph Generating Dependency approach

sHINER[12] is a tool that can be used to recommend missing links among graph entities based on a predefined set of Graph Generating Dependencies (GDDs) [4]. GGDs represent a novel class of dependencies for property graphs, extending the concept of tuple-generating dependencies to graph data. GGDs express constraints between two potentially different graph patterns, enforcing relationships based on both data properties and graph topology. This class of dependencies addresses the need for advanced data management techniques in graph-structured data, facilitating tasks such as data integration, data quality assurance, and query optimisation.

In the context of SciLake, we used sHINER to link entities in Scientific KGs (SKGs). An indicative example of this was a GGD used for link recommendation for the OpenAIRE Graph to suggest links between results and projects, in case that the result has author(s) from an organisation that participates in the respective project. Figure 4.2.2.1 illustrates the respective GGD: its main goal is to connect each project to all results the organisation has previously worked on. By linking the project to the results we can easily visualise how different projects relate to the same result and consequently use these new links to make queries automatically. The same approach is expected to be used to identify other types of missing links between entities of the OpenAIRE Graph and/or the domain-specific SKGs.

[12] github.com/smartdatalake/gcore-spark-ggd

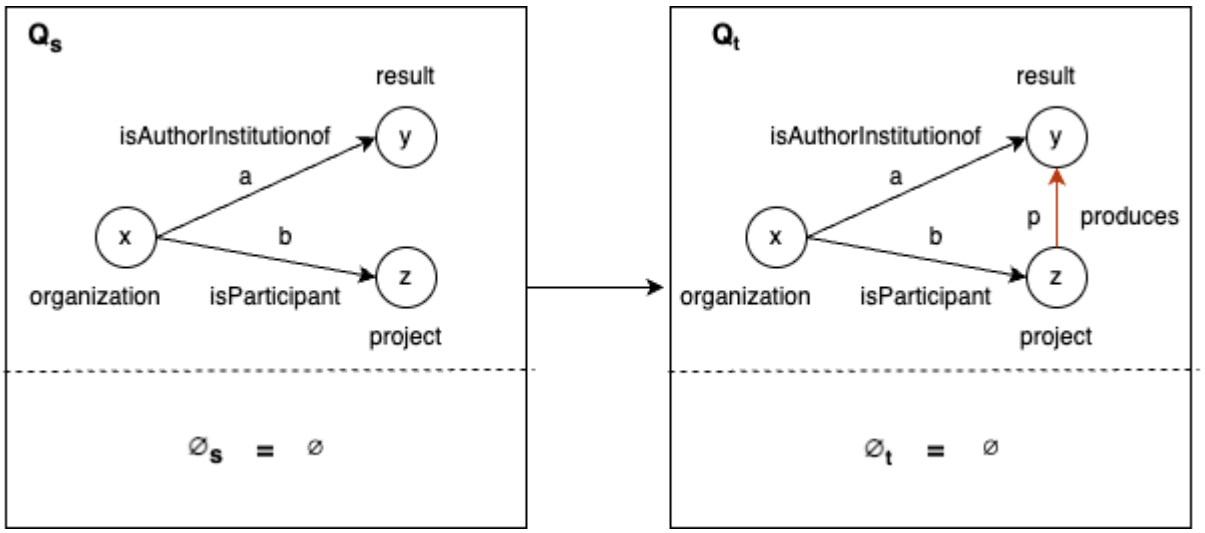

Figure 4.2.2.1: *Illustration of the GGD used to suggest links between projects and results in the OpenAIRE Graph based on the organisations involved in the results.*

4.2.3. SciNeM: the metapath approach

SciNeM[13] [5], a powerful Knowledge Graph (KG) analysis engine, focuses on a particular type of KG analysis, called "metapath-based analysis". This type of analysis is capable of exploiting the structure of the graph to extract latent knowledge (e.g., missing links) that is encoded inside it. To better understand the respective concept, we need to provide a quick introduction on the main terms involved.

KGs offer an intuitive and generic model to encapsulate complex semantic information via different types of nodes and edges, which both have internal structure including a set of properties. The next figure illustrates an example graph with its schema (that can be explicitly or implicitly defined). It consists of nodes representing papers (P), authors (A), venues (V), and topics (T) and (bidirectional) edges of three types: authors – papers (AP / PA), papers – topics (PT / TP), and papers – venues (PV / VP).

[13] https://github.com/athenarc/SciNeM-workflows & https://github.com/athenarc/SciNeM

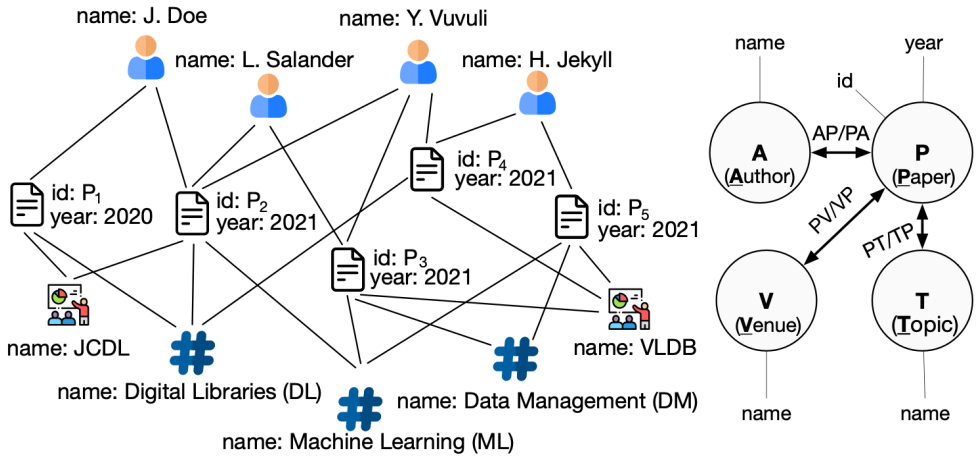

Figure 4.2.3.1: *A KG example.*

Indirect relationships between entities are implicitly encoded by paths in the graph. All paths that correspond to the same sequence of node and edge types (i.e., the same "metapath" [6]) encode latent relationships of the same interpretation between the starting and ending nodes. In the above example, the metapath $\langle APTPA \rangle$ relates authors that have published papers on the same topic (e.g., both "J. Doe'' and "H. Jekyll'' have papers about "DL"). Notice that each metapath corresponds to a path on the schema of the graph. Metapaths are instrumental for KG analysis: the metapath-based connectivity can be used to define node similarity measures or to rank nodes based on their centrality in a metapath-defined graph/network. In our example, using $\langle TPT \rangle$ to perform a similarity join could reveal that topics "ML" and "DL" are very related since they are involved in two common papers. To further elaborate the analysis, it is often useful to apply property-based constraints to metapaths (e.g., to consider only metapath instances involving papers published later than "2020"). It should be noted that the first step in any metapath-based analysis is to generate a transformed graph that contains edges between all pairs of nodes connected with one or more paths of a specified type.

Based on the previous discussion, it is now easier to understand the various components of the SciNeM tool. The following figure presents the high-level architecture of SciNeM.

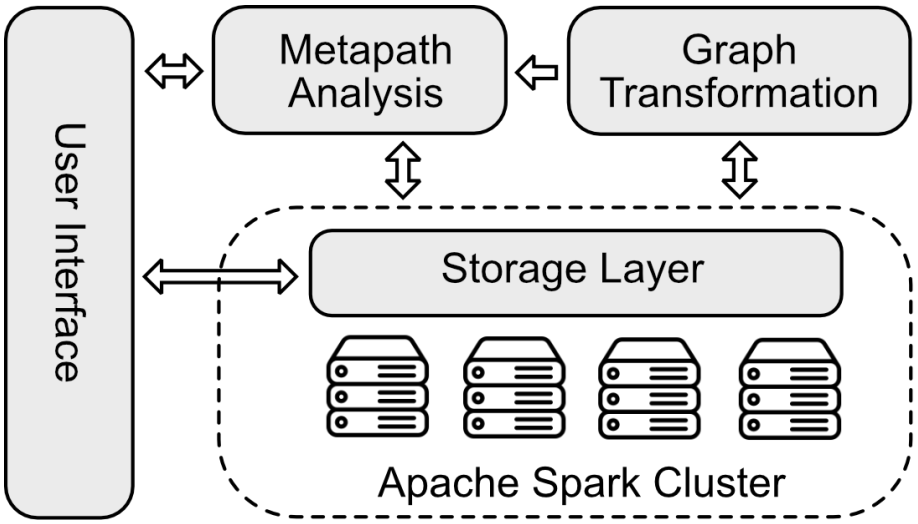

Figure 4.2.3.2: *Overview of the SciNeM architecture.*

SciNeM is a distributed analysis tool that can be deployed on an Apache Spark[14] computational cluster. Its storage layer, which stores all KG data, utilises a Hadoop Distributed File System (HDFS) hosted on the storage media of the underlying computational cluster. Each KG comprises a set of files, including:

- a ***schema file*** (compatible with Cytoscape's[15] Elements JSON format) that describes the node types of the KG and the types of relationships among them,
- ***node files*** in TSV format containing data attributes for the nodes of each type, and
- ***relationship files*** that define the network edges.

User-created KGs, which are compatible with the previous format, can be uploaded to SciNeM's distributed storage layer via its web-based user interface (UI).

The core of the SciNeM tool is its graph transformation component. This component implements the common preprocessing step that converts the initial heterogeneous graph into a transformed version that contains edges between all pairs of nodes connected with one or more paths of a specified type. The type of paths is given by the user and it essentially captures latent relationships with particular semantics. The transformed graph is later used to apply a desired type of analysis on top. In the context of the SciLake project, our interest was on missing links recommendations. We started working on some relationship types of the OpenAIRE Graph that were deemed interesting, where there was space for improvements in the coverage. For instance, we explored providing recommendations for linking publications to projects or communities by analysing the similarity between new publications and those

[14] spark.apache.org
[15] cytoscape.org

already associated with the target project or community. To this end, we used metapath-based similarities with interesting semantics, following a similar approach like the one we have used in the past for another work [7]. The preliminary results are interesting but more thorough experiments are needed. In addition, we plan to follow a similar approach for the other SKGs of SciLake, after consultation with the pilot representatives.

It is worth mentioning that, since the calculation of the transformed graph is a computationally intensive task, special care was taken for the efficient implementation of this component. In the original version, the core of the transformation is calculated using matrix multiplication between the adjacency matrices defined by the relations of the given metapath. Specifically, our approach is based on the work in [8] but extends it by utilising sparse matrix representations. Since the order of multiplications significantly affects the performance of the whole processing, we adopt a dynamic programming approach that estimates the optimal ordering taking into consideration the computational cost of sparse matrix multiplications introduced in [9]. This modification offers significant speedups, compared to the baseline, in many cases. In addition, since the implementation of this component utilises Apache Spark, it takes advantage of parallel and distributed computing.

In the context of the project, we have also investigated a different approach, called Atrapos [10], which efficiently detects, in real time, frequent metapath overlaps within a sequence of queries. To do so, it uses: (i) sparse matrix representations, which leverage the fact that matrices involved in metapath computations are largely sparse, especially for constrained metapaths; and (ii) a tailored data structure with a customised caching policy to materialise reoccurring intermediate results. Our experimental evaluation demonstrated that Atrapos outperforms traditional, single-query approaches in the evaluation of metapath query workloads, while it is considerably faster than baseline approaches. The detailed methodology and experiments can be found in the original publication.

Paper(s):

- SciNeM: A Scalable Data Science Tool for Heterogeneous Network Mining
- Atrapos: Real-time Evaluation of Metapath Query Workloads.

Code repo(s):

https://github.com/athenarc/SciNeM-workflows
https://github.com/athenarc/SciNeM
https://github.com/schatzopoulos/ATRAPOS

4.3. Article segmentation for multilingual articles

4.3.1. Article Segmentation

The objective of this tool is the recognition of the structure of scientific documents, specially identifying its constituent parts (sections and titles). Apart from the identification of parts of the document, a second functionality is the classification of the parts in predefined section/part types—e.g, the conclusions in a scientific paper could be named in many ways, such as 'Conclusions', 'Conclusions and Future Work', 'Discussion', etc.

This functionality is depicted in Figure 4.3.1.1.

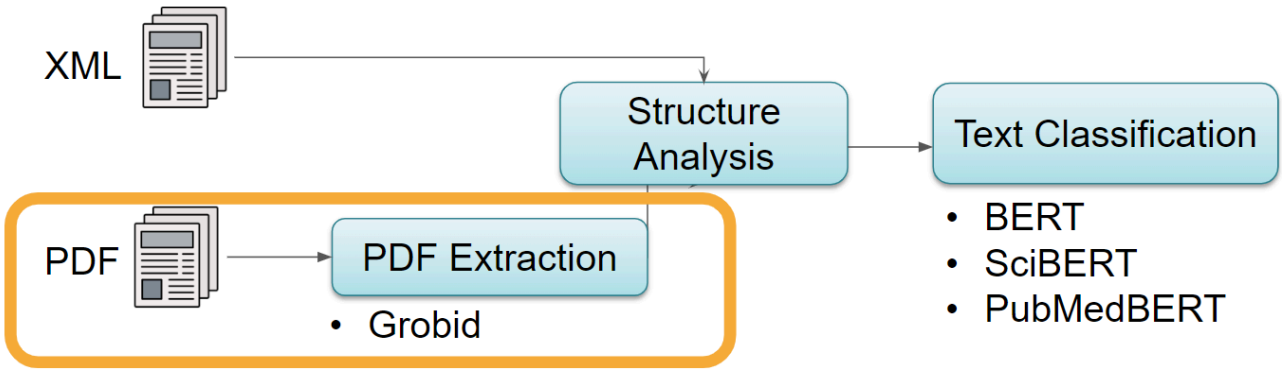

Figure 4.3.1.1: *Article Segmentation for scientific articles.*

4.3.1.1. Scientific Article Structure Detection

The functionality of the service is the processing of PDF documents to identify structural components, such as headings, texts, tables, figures, references, etc. This service is based on GROBID[16], an external tool that helps in the identification of structural components in scientific articles. The service has been implemented in Java because GROBID is also implemented in Java, and therefore the integration of both tools is easier.

The following functionalities are available in GROBID: header extraction, references extraction, citation contexts recognition and resolution of the full bibliographical references of the article, full text extraction and structuring from PDF articles, PDF coordinates for extracted information, parsing of references in isolation, parsing of names, parsing of affiliation and address blocks, parsing of dates, consolidation/resolution of the extracted bibliographical references and extraction and parsing of patent and non-patent references in patent publications. In a complete PDF processing, GROBID manages 55 final labels used to build relatively fine-grained structures, from traditional publication metadata (title, author first/last/middle names, affiliation types, detailed address, journal, volume, issue, pages, DOI, PMID, etc.) to full text structures (section title, paragraph, reference markers, head/foot notes, figure captions, etc.).

The document structure recognition (DSR) component integrates GROBID using the available JAVA API, and it offers three output formats:

- XML format using TEI schema.
- JSON format using TEI schema.
- RDF format mapped from TEI.

Apart from that, the REST API access has been implemented using the python library Flask. The tool is provided as a docker container (URLs are provided in Table 4.3.1.3) using Linux as the base operating system. The container is provided so that it can be utilised completely independently, without any special configuration needed.

[16] github.com/kermitt2/grobid

4.3.1.2. Scientific Article Section Classification

We train a text classifier to automatically classify each part of the text to a predefined section class.

Text Classification Data

Data for text classification is obtained from the PubMed dataset. There are 164,195 unique section titles (STs) in PubMed, which imply the section content and describe the common topic for its sub-sentences. Similar STs like "Conclusion" and "Conclusions" can be grouped into one typical class of "Conclusion". We manually predefine a dictionary of 8 typical section classes and the corresponding in-class STs.

The classes are: introduction, background (i.e., background, review and related work), case (i.e., case reports), method, result, discussion, conclusion and additional information (such as conflicts of interest, financial support and acknowledgements). Each in-class section text is extracted from the original dataset and assigned with the corresponding section class label. In this way, we generate 490,452 samples for multi-class text classification. For each section class, 20% of the samples are randomly selected as the test data, the rest samples are used as the training data. The detailed statistics of the data are summarised in Table 4.3.1.1.

Section Class	# samples	# train / # test
0. introduction	98,188	78,550 / 19,638
1. background	6,731	5,384 / 1,347
2. case	34,619	27,695 / 6,924
3. method	107,741	86,192 / 21,549
4. result	63,635	50,908 / 12,727
5. discussion	92,202	73,761 / 18,441
6. conclusion	70,240	56,192 / 14,048
7. additional	17,096	13,676 / 3,420
total	**490,452**	**392,358 / 98,094**

Table 4.3.1.1: *Data for Text Classification.*

Text Classification Model

Our Text Classification Model consists of a pre-trained Transformer Language Model (TLM) for text encoding and a sigmoid classifier for multi-class classification. The entire model is trained on the training data for 3 epochs with a training batch of 16. In experiments, the effect of three different TLMs are investigated: BERT[17], SciBERT[18] and PubMedBERT[19].

Evaluation Results

The evaluation results of Text Classification on the test data are summarised in Table 4.3.1.2. Demonstrated are three classification models based on the three TLMs respectively and the corresponding runtime (for training and evaluation) as well as the classification accuracy. Comparing the first two models, it is clearly observed that the classification accuracy is increased by using SciBERT which was pre-trained in the scientific domain. Using PubMedBERT, which was pre-trained especially on PubMed and thus captures in-depth understanding in biomedical sciences, the classification accuracy is further increased by 0.4.

Model	Runtime	Accuracy
CLS_BERT	33h	94.58
CLS_SciBERT	21h	94.72
CLS_PubMerBERT	21h	95.12

Table 4.3.1.2: *Evaluation Results of Text Classification.*

Next steps

We will further develop the Article Segmentation tool, including the multilingual functionality to the Section Classification submodule. We are currently working on offering a new linked data output format where the structure and classification information of the scientific publication is encoded using existing and, if needed, self-defined ontologies.

[17] huggingface.co/google-bert/bert-base-uncased
[18] huggingface.co/allenai/scibert_scivocab_uncased
[19] huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext

The next table summarises the most important resources (publications, code repositories, etc) related to the DSR component.

Code repo(s): https://gitlab.com/dfki-scilake/dsr **Docker URL:** registry.gitlab.com/dfki-scilake/dsr:d41

Table 4.3.1.3: *Links to code repositories and documentation of the DSR component.*

4.3.2. Automatic translation

For the automatic translation service, we have developed a suite of Neural Machine Translation (NMT) models. These models are specifically tailored to address the challenges posed by the translation of scientific texts across the four pilots used in SciLake: Cancer Research, Energy Research, Neuroscience, and Transportation Research.

We began by constructing comprehensive parallel and monolingual corpora targeting the language pairs of Spanish-English, French-English, and Portuguese-English. These language pairs have been selected based on the amount of available open, high-quality scientific data. The corpora were meticulously curated to include general scientific content as well as specialised domains based on the four pilots.

Utilising these corpora, we fine-tuned open-source, general-purpose NMT systems to enhance their performance significantly, ensuring translations are not only fluent but also precise and faithful to the original scientific content. To this end, we chose to use selected pre-trained models from OPUS-MT [11] as our baseline systems. We chose these models because they are all based on state-of-the-art transformer-based NMT architectures, and they were trained on freely available parallel corpora from the OPUS[20] bitext repository. Even though these models are not considered to produce the best quality for the selected language pairs, they provide a robust initial performance across a variety of language pairs. We have released three translation models, one for each language pair, on the Hugging Face platform:

- **French-English:** huggingface.co/ilsp/opus-mt-big-fr-en_ct2_ft-SciLake
- **Portuguese-English:** huggingface.co/ilsp/opus-mt-pt-en_ct2_ft-SciLake
- **Spanish-English:** huggingface.co/ilsp/opus-mt-big-es-en_ct2_ft-SciLake

[20] opus.nlpl.eu/

In order to determine the improvement on translation quality for each specific domain of interest, we perform evaluation using 4 domain specific test sets, as well as a General Scientific set of test sentences. To that end, the most widely-used benchmarks for the evaluation of automatic translations have been chosen, these being BLEU, chrF2++, and COMET. Our evaluation results for each model can be found in the appropriate model link (see above).

The produced models can be easily provided as API endpoints for the SciLake translation service. Our next steps are to create a translation pipeline, so that the translation models will be available as services in the SciLake project for the translation of various types of documents. We are also planning to experiment with document-level translation models so as to better handle longer text and inter-sentential dependencies.

4.4. SciNoBo CA tool: Citation-context assisted replication assessment

The Citation-context Assisted Replication Assessment component utilises advanced NLP techniques to analyse scientific citations. Powered by the SciNoBo Citance Analysis (CA) tool[21], it determines the Intent, Polarity, and Semantics of citation mentions (citances) within scientific literature.

The primary objective of the Citation-context Assisted Replication Assessment is to provide a nuanced analysis of citations, aiding in the replication and validation of scientific findings. Specifically, it addresses the following research questions:

- What is the purpose or intent of the citation?
- What stance do the authors take towards the cited work?
- What aspect of the cited work is being referenced?

[21] github.com/iNoBo/scinobo-citance-analysis

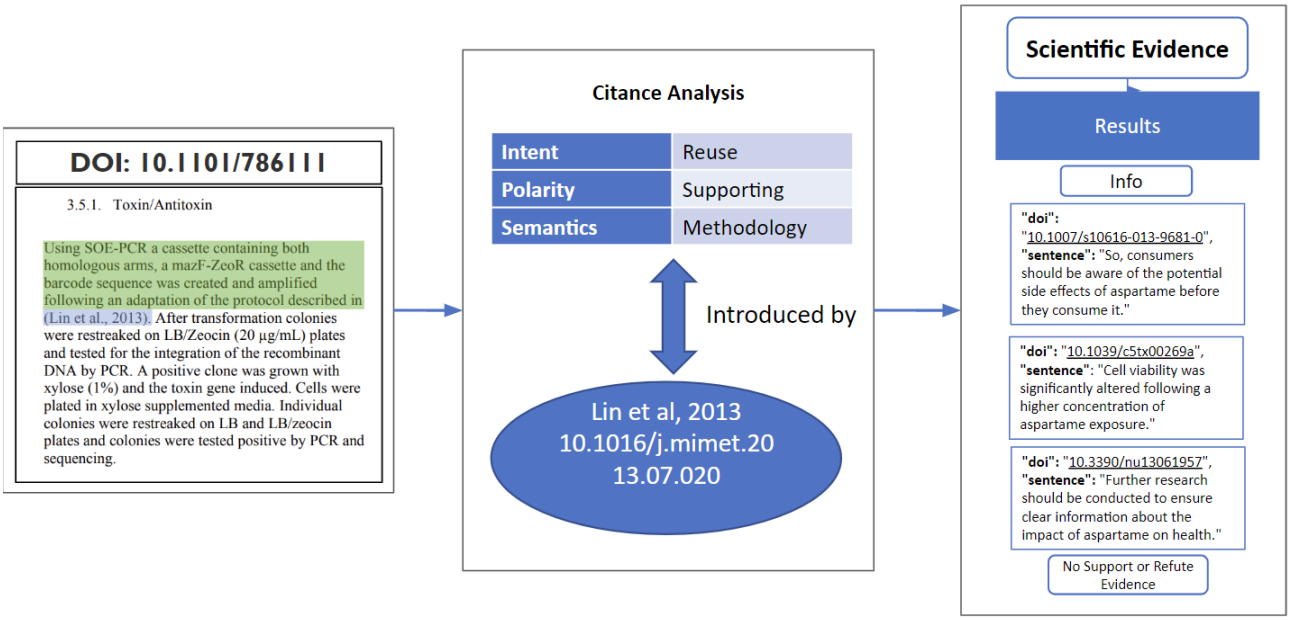

Figure 4.4.1: *Citance analysis example.*

4.4.1. Methodology

The input data for this analysis consists of scientific citation mentions (citances) extracted from publication PDFs using GROBID. The analysis involves three main components: **Intent Analysis**, **Polarity Analysis**, and **Semantics Analysis**. Each component serves a distinct purpose in understanding the nature and context of the citances.

Intent Analysis aims to identify the purpose of the citation within the text. This involves classifying citances into one of three categories:

- Generic references, where the citation is a general acknowledgment of the cited work.
- Reuse instances, where the authors incorporate an aspect of the cited work into their own research.
- Comparison scenarios, where the authors draw comparisons between their work and the cited work.

Polarity Analysis determines the stance or attitude of the authors towards the cited work. This is classified into three categories:

- Supporting, where the authors endorse or support an aspect of the cited work.
- Neutral, where the authors neither support nor refute the cited work.
- Refuting, where the authors challenge or refute an aspect of the cited work.

Semantics Analysis classifies the content or meaning of the citation. The citances are categorised based on their reference to specific elements of the cited work:

- Claim, where the authors refer to a specific claim made in the cited work.
- Methodology, where the citation pertains to the methods or approaches used in the cited work.
- Results, where the authors discuss the findings or results presented in the cited work.
- Artefact, where the citation references a specific research artefact such as a dataset, software, or model.

In the current state, the system utilises a fine-tuned version of the Flan T5 Base model[22], a powerful text-to-text transformer model designed for diverse language processing tasks. This model is fine-tuned using the Low-Rank Adaptation (LoRA) method [12], which facilitates efficient adaptation of pre-trained language models to specific tasks with limited data. The fine-tuning process leverages a very small synthetic dataset that includes examples representing all possible combinations of Intent, Polarity, and Semantics.

An instruction-based Question Answering (QA) setting is used for this classification task, similar to [13] and the SciNoBo Research Artefact Analysis (RAA) tool in Section 3.1.1. In this setup, different prompts are employed to classify the intent, polarity, and semantics of citances. For the classification, the sentence of the citances is given as input to the model, along with its paragraph as context for better understanding.

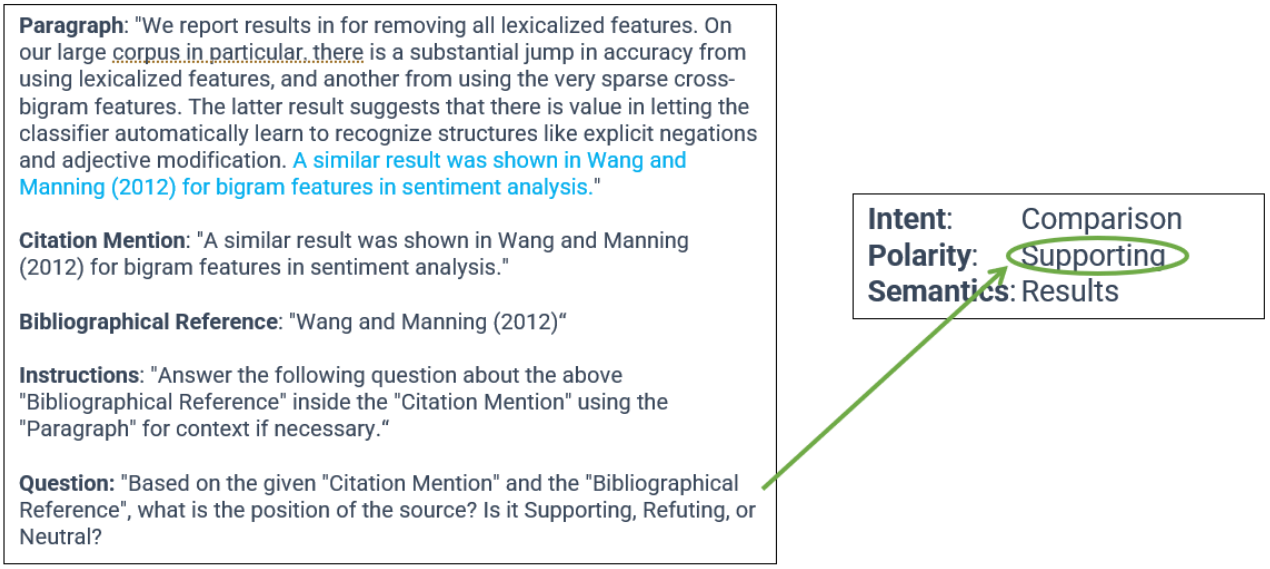

Figure 4.4.1.1: *Example of the instruction-based Question Answering setting.*

[22] huggingface.co/google/flan-t5-base

4.4.2. Next steps

The next steps for enhancing the Citation-context Assisted Replication Assessment component and the SciNoBo Citance Analysis (CA) tool include:

1. Transition to state-of-the-art LLMs such as LLaMA 3. These models offer better performance compared to the current Flan T5 model and will enable more accurate analysis of citances.
2. Use the current system as a bootstrap to gather real-world citances. Annotate and curate these citances to increase the dataset size, which will improve the model's training and performance.
3. Inference with the tool on publications from the SciLake pilots. This real-world testing will provide valuable insights into the tool's effectiveness and areas for improvement.
4. Collect feedback from the SciLake pilots and evaluate the tool's performance across different disciplines. This will help determine if the tool works well in various fields or if it needs adjustments for specific disciplines.
5. Add more classes for Intent and Semantics based on feedback from the SciLake pilots. For example, new categories like the "Expansion Intent" could capture more specific aspects of citances, enhancing the tool's analytical capabilities.

Code repo(s):
https://github.com/iNoBo/scinobo-citance-analysis

Docker hub:
https://hub.docker.com/repository/docker/intelligencenoborders/scinobo-citance-analysis/general

HF Space (Demo):
https://huggingface.co/spaces/iNoBo/scinobo-citance-analysis

4.5. Reproducibility and replicability badges

In this initial version of the service, the implementation of the above will be carried out in the form of badges of reproducibility and replicability integrated into the UI of the Smart Impact-driven Discovery Service (see D3.1). This enables users to readily identify research outputs with detailed methodological information, as well as those reporting findings that have been subsequently verified or, to the contrary, may have potential flaws. In a future iteration, accompanied by guidelines and best-practice recommendations to improve reproducibility, the service will make it possible for researchers to examine their own outputs

under the same light, in order to provide an overall assessment of their status and suggest possible areas of improvement.

5. Conclusions

We have presented the initial version of the Smart Reproducibility Assistance Service, which incorporates methodologies such as text mining, link prediction, and citation context analysis to enhance the Scientific Knowledge Graphs within the Scientific Lake. Leveraging the contents of the underlying SKGs, the components of the service come together to guide researchers in the discovery of credible research outputs in the respective SKGs, based on an assessment of their reproducibility and replicability. This development aims to provide researchers with tools to better understand and pursue reproducibility in scientific findings.

Future work will focus on expanding the capabilities of the different components, collaborating with pilot representatives to meet their domain-specific needs, and integrating additional functionalities to further support reproducibility assessment. Specifically, we plan to: i) improve the accuracy and scope of the SciNoBo RAA tool to recognise a broader range of research objects across more domains; ii) optimise sHINER and SciNeM methodologies to provide more accurate and contextually relevant link recommendations; iii) implement more sophisticated algorithms for detecting and classifying sections of scientific articles, including support for additional languages and complex document structures; iv) enhance the SciNoBo CA tool with more robust models for evaluating replication based on diverse citation contexts; vi) develop and train new NER models to identify additional entities based on the specific needs of our pilot domains (this will involve leveraging existing annotated datasets and, if necessary, generating new annotated datasets in collaboration with the pilot projects); and vi) finalise the criteria and implementation for badges that signify reproducibility and replicability, providing clear and recognisable indicators of research quality. The ongoing improvements are expected to contribute to higher standards of transparency and reliability in scientific research.

6. References

[1] Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature, 533(7604).

[2] Arie Cattan, Sophie Johnson, Daniel S Weld, Ido Dagan, Iz Beltagy, Doug Downey, & Tom Hope (2021). SciCo: Hierarchical Cross-Document Coreference for Scientific Concepts. In 3rd Conference on Automated Knowledge Base Construction.

[3] Luo, L., Wei, C. H., Lai, P. T., Leaman, R., Chen, Q., & Lu, Z. (2023). AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning. Bioinformatics, 39(5), btad310.

[4] Shimomura, L. C., Fletcher, G., & Yakovets, N. (2020, October). Ggds: Graph generating dependencies. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 2217-2220).

[5] Serafeim Chatzopoulos, Thanasis Vergoulis, Panagiotis Deligiannis, Dimitrios Skoutas, Theodore Dalamagas, Christos Tryfonopoulos. SciNeM: A Scalable Data Science Tool for Heterogeneous Network Mining. EDBT 2021: 654-657.

[6] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, and Tianyi Wu. 2011. Path- Sim: Meta Path-Based Top-k Similarity Search in Heterogeneous Information Networks. PVLDB 4, 11 (2011), 992–1003.

[7] Serafeim Chatzopoulos, Thanasis Vergoulis, Theodore Dalamagas, Christos Tryfonopoulos. VeTo+: improved expert set expansion in academia. Int. J. Digit. Libr. 23(1): 57-75 (2022)

[8] C. Shi, Y. Li, P. S. Yu, and B. Wu. 2016. Constrained-meta-path-based ranking in heterogeneous information network. Knowl. Inf. Syst. 49, 2 (2016), 719–747.

[9] D. Kernert, F. Köhler, and W. Lehner. 2015. SpMacho - Optimizing Sparse Linear Algebra Expressions with Probabilistic Density Estimation. In Proc. of the 18th International Conference on Extending Database Technology, EDBT 2015, Brussels, Belgium. 289–300

[10] Serafeim Chatzopoulos, Thanasis Vergoulis, Dimitrios Skoutas, Theodore Dalamagas, Christos Tryfonopoulos, Panagiotis Karras. Atrapos: Real-time Evaluation of Metapath Query Workloads. WWW 2023: 2487-2498

[11] Tiedemann, J., & Thottingal, S. (2020, November). OPUS-MT–building open translation services for the world. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (pp. 479-480).

[12] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

[13] Stavropoulos, P., Lyris, I., Manola, N., Grypari, I., & Papageorgiou, H. (2023, December). Empowering Knowledge Discovery from Scientific Literature: A novel approach to Research Artifact Analysis. In Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023) (pp. 37-53).