

Sci!lake

Scientific Lake

DOI 10.5281/zenodo.12806147

Deliverable D3.1: Initial version of the smart impact-driven discovery service

Due Date of Deliverable	30/06/2024
Actual Submission Date	30/06/2024
Work Package	WP3
Tasks	T3.1 T3.2 T3.3 T3.4 T3.5
Type	OTHER
Approval Status	Submitted
Version	v1.0
Number of Pages	41
The information in this document reflects only the author's views and the European Commission is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.	

Abstract

In this deliverable report, we describe the beta release of a smart knowledge discovery service, which has been developed in the context of the SciLake project and leverages SciLake's Scientific Knowledge Graphs (SKGs) and the various technologies provided by the Scientific Lake tier of SciLake's architecture. The service is designed to leverage the contents of the Scientific Lake to calculate indicators of scientific impact for research products and offer, based on them, useful functionalities to researchers so that they could (a) prioritise their reading of the relevant literature when searching for a specific subject of interest (b) uncover latent knowledge that could be instrumental in accelerating their research conclusions, and (c) identify emerging trends related to research topics of interest. The report provides references to the respective open code bases and elaborates on the technical details related to the development of the various components.



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
0.1	9/5/2024	Agreement on structure	Thanasis Vergoulis
0.1	27/5/2024	First Draft	Thanasis Vergoulis
0.2	25/6/2024	Intermediate version	Thanasis Vergoulis
0.3	25/6/2024	Peer review start	Thanasis Vergoulis
0.4	29/6/2024	Peer review comments addressed	Thanasis Vergoulis
1.0	30/6/2024	Final Version after proofreading	Thanasis Vergoulis

Author List

ORGANISATION	NAME	CONTACT INFORMATION
ATHENA RC	Thanasis Vergoulis	vergoulis@athenarc.gr
ATHENA RC	Serafeim Chatzopoulos	schatz@athenarc.gr
ATHENA RC	Sotiris Kotitsas	sotiris.kotitsas@athenarc.gr

Contributor List

ORGANISATION	NAME	CONTACT INFORMATION
CNR	Miriam Baglioni	miriam.baglioni@isti.cnr.it
SIRIS	Pablo Accuosto	pablo.accuosto@sirisacademic.com

Table of Contents

1. Executive Summary	6
2. Introduction	7
3. Design	9
3.1. Main requirements	9
3.2. Design approach	11
3.3. Integration into SciLake ecosystem	11
4. Implementation	12
4.1. Field extraction component	13
4.1.1. FoS Taxonomy	13
4.1.2. FoS Classifier	17
4.1.3. FoS Taxonomy Mapper	20
4.1.4. SDG Classifier	22
4.2. Citation-based impact analysis component	24
4.3. Topic analysis component	29
4.4. Knowledge discovery portal	31
5. Demonstrated Use Case	36
6. Conclusion	39
7. References	40

DRAFT

List of Tables

Table 1: Example of a Level 4 FoS field under the domain of AI.

Table 2: Example of a cluster of NCs of a Level 4 FoS field.

Table 3: Links to the repository and publications related to the Field of Science Taxonomy.

Table 4: Links to code, docker and demo repositories of the Field of Science Classifier.

Table 5: Concatenated paths from Level 2 up to Level 6.

Table 6: Links to code, docker and demo repositories of the Field of Science Taxonomy Mapper.

Table 7: Links to code, docker and demo repositories of the SDG Classifier.

Table 8: Links to code repositories and documentation of the citation-based impact analysis component.

Table 9: Links to code repositories and documentation of the topic analysis component.

Table 10: Links to code repositories and documentation of the knowledge discovery portal component.

List of Figures

Figure 1: A snapshot of the Field of Science Taxonomy.

Figure 2: A snapshot of the multilayer graph of the FoS classifier.

Figure 3: FoS classifier process.

Figure 4: Pipeline of the SDG algorithm.

Figure 5: A snapshot of the BIP! Spaces user interface displaying topics.

Figure 6: A snapshot of the BIP! Spaces user interface displaying domain-specific annotations coming from an SKG.

Figure 7: Details for a particular annotation in BIP! Spaces UI.

Figure 8: Impact indicators in the BIP! Spaces UI.

Figure 9: Screenshot from the BIP! Space administration UI showing the current annotations being determined for the cancer research space.

Abbreviation List

- **FoS:** Fields of Science
- **HF:** Hugging Face
- **KG:** Knowledge Graph
- **LLMs:** Large Language Models
- **NC:** Nominal Chunk
- **POS:** Part of speech
- **SDGs:** Sustainable Development Goals

- **SKG:** Scientific Knowledge Graph
- **UI:** User Interface

1. Executive Summary

In today's fast-paced world of scientific research, the sheer volume of available research outputs—such as publications, datasets, and software—has a dual impact on researchers. On one hand, the abundance of valuable information provides a rich foundation for their work. On the other hand, navigating this vast knowledge space can be overwhelming, particularly given the varying quality of research works available. Contrary to the past, today's researchers face the challenge not of a lack of information, but of efficiently discovering relevant and high-quality knowledge.

Building services that facilitate scientific knowledge discovery is crucial for accelerating innovation and progress. By providing researchers with advanced tools to seamlessly search, filter, and access the most valuable and relevant to their research publications, datasets, and findings, we empower them to spend more time on actual research and less on navigating through information overload. Finally, as a side effect, such services can democratise access to information, ensuring that researchers from diverse backgrounds and institutions have equal opportunities to contribute to and benefit from the collective pool of scientific knowledge.

The concept of the Scientific Lake, which is being introduced by the SciLake project, offers a foundation of cutting-edge technologies that can significantly simplify the development of added-value services aiming to facilitate scientific knowledge discovery. At the core of this concept are Scientific Knowledge Graphs (SKGs), which provide an efficient means of organising, querying, and accessing both domain-agnostic and domain-specific knowledge. This facilitates the uncovering of hidden connections between research-related entities and highlights emerging trends, making them invaluable tools for advancing scientific discovery and enabling informed decisions for the work of researchers. Finally, SciLake also puts effort on data models and specifications that can make the SKGs more interoperable to each other further simplifying the creation of added-value services on top of them.

In this report, we describe the beta release of a smart knowledge discovery service, which has been developed in the context of the SciLake project and leverages SciLake's SKGs and the various technologies provided by the Scientific Lake tier of SciLake's architecture. The service is designed to leverage the contents of the Scientific Lake to calculate indicators of scientific impact for research products and offer, based on them, useful functionalities to researchers

so that they could (a) prioritise their reading of the relevant literature when searching for a specific subject of interest (b) uncover latent knowledge that could be instrumental in accelerating their research conclusions, and (c) identify emerging trends related to research topics of interest. The service is also configurable so that it can be tailored based on the needs and the particularities of different scientific domains. In the context of the project, different instances of the service are configured based on the domains of the project pilots and these instances are used to demonstrate the service and evaluate its value, as well the value of the underlying Scientific Lake.

In the following sections, we provide a comprehensive overview of the SciLake smart knowledge discovery service. Section 2 delves into the problem that SciLake aims to solve with this service, elaborating on the background, the motivation of the respective work, and the associated challenges. Section 3 outlines the design process followed for the service. In Section 4, we offer implementation details for all the major components of the service and in Section 5, we present use cases that demonstrate its value. Finally, Section 6, summarises the report and briefly discusses the next steps in further developing the service.

2. Introduction

In recent years, the volume of scientific and research output has surged dramatically. This explosion in scientific content is driven by several factors, including the increasing number of researchers entering various fields [1] and the pervasive "publish or perish" culture that pressures academics to continually produce and publish new research. On one hand, this proliferation of research is beneficial for researchers as it results in a vast repository of knowledge that researchers can draw upon to advance their own studies. The extensive range of available content fosters innovation, cross-disciplinary discoveries, and rapid advancements in science and technology.

However, this rapid increase in research output also presents significant challenges. The sheer volume of published work makes it increasingly difficult for researchers to identify valuable and relevant studies amidst this overload of information. Various studies and reports [2] have highlighted that the accelerated pace of research production can lead to a decline in quality, with some research works lacking rigorous peer review or robust methodologies. Consequently, it becomes more challenging for researchers to discern high-quality information potentially hindering scientific progress. This dual-edged nature of the burgeoning research landscape underscores the need for advanced tools and services that can help researchers efficiently sift through vast amounts of data to find the most pertinent and high-quality information.

To make matters worse, current scientific knowledge discovery services tend to focus predominantly on publications, neglecting the diverse and heterogeneous nature of modern

research output. It is increasingly recognized that valuable research contributions come in many forms beyond traditional papers, including research data, software, peer reviews, and more. This narrow focus on publications creates significant challenges for researchers who need comprehensive access to all relevant resources to advance their work. Addressing this issue is crucial because each type of research output provides unique insights and plays a vital role in the scientific process. Research data can offer detailed experimental results and raw information that underpins published findings, while software tools and algorithms developed during research projects are essential for replicating studies and building upon previous work. Peer reviews and other forms of scholarly communication provide critical evaluations and discussions that enrich the understanding of published research. Therefore, developing knowledge discovery services that encompass a wider array of resources is essential for fostering a more holistic and effective research environment. By doing so, we can ensure that researchers have access to a complete spectrum of scientific contributions, facilitating better-informed decisions, fostering innovation, and ultimately advancing the frontiers of knowledge more efficiently.

In our opinion, the Scientific Lake concept introduced by the SciLake project holds significant potential for paving the way to address these challenges. As detailed in D2.1, this concept offers a suite of valuable components that enable domain experts to better extract and organise domain-specific knowledge, facilitating the development of added-value services for scientific discovery. At the core of this concept are various Scientific Knowledge Graphs (SKGs), which provide efficient and effective ways to query and explore the encoded knowledge, helping service developers in revealing valuable insights.

Given the importance of knowledge discovery activities for science, the aforementioned challenges, and the potential of the Scientific Lake concept to help address these challenges we have chosen to build a smart knowledge discovery service on top of the Scientific Lake that we are building in the context of the project. By leveraging the comprehensive capabilities of SKGs, this service can significantly enhance the ability of researchers to navigate the vast and diverse landscape of scientific outputs in the scientific domains of interest. The SciLake project's approach ensures that the diverse nature of research contributions, including datasets and research software is integrated and accessible, thus providing a more holistic and enriched platform for scientific discovery.

3. Design

In this section, we quickly discuss the main requirements of the service, the process followed during its design, and how it is integrated into the SciLake ecosystem. More details about the design process can be found in Deliverable D1.2 (“Initial integrated system”).

3.1. Main requirements

Our aim for SciLake’s smart knowledge discovery service is to provide researchers with a powerful tool to navigate the vast knowledge space of their domain while also offering valuable insights for their research activities. More specifically, the service should be able to support the following main functionalities:

- Offer advanced keyword search functionalities for research products and domain-specific entities, providing multiple ways of presenting, filtering, and ordering the search results.
- Offer integration with domain-specific SKGs, exploiting their contents to present the keyword search results together with valuable enrichments coming from the underlying SKGs in an intuitive and insightful way.
- Calculate advanced citation-based impact indicators for research products and leverage them in (a) ordering, filtering, and/or organising the search results to facilitate the identification of valuable research products and new knowledge discovery, (b) providing analytics and visualisations that can help in identifying and monitoring trends in scientific topics and their evolution.
- Implement impact propagation and aggregation techniques to calculate enhanced citation-based indicators for various types of research products beyond publications, such as datasets, as well as for related domain-specific entities, investigating also the enhancement of citation links with other types of acknowledgement links between research products (e.g., in-text mentions) or semantics (e.g., leveraging citation intent classification).
- Calculate classifications of research products in scientific fields and domains based on their metadata and contents and use these classifications to (a) provide useful enrichments for the search results, (b) provide insights about the level of impact of research products into particular scientific fields/domains, and (c) give valuable input to the components providing analytics and visualisations for topic trends and evolution.

Based on the previous, the service is expected to have the following distinct components to support the aforementioned functionalities:

- Field-extraction component
- Citation-based impact analysis component
- Topic analysis component
- Knowledge discovery portal

In addition, recognizing that each scientific domain has unique requirements for knowledge discovery, encompassing, for instance, distinct domain entities and research methodologies, it became evident that we should design the service with adaptability in mind. It was crucial to ensure that the service could be tailored to accommodate the diverse needs of each pilot domain effectively. As such, the service should be configurable to meet the following requirements:

- The service should support multiple ways to order search results and should be possible for service instance administrators to configure which are the ordering options provided to the end-user for the particular instance as well as which is the default option for results ordering for the instance.
- The service should be configurable by service instance configurations regarding the type of research products and/or domain entities that will be presented to the end-user during the knowledge discovery process.
- The service should be able to present a particular selection of enrichments for the search results based on the SKGs connected to the particular service instance and this selection should be configurable by the service instance administrators.
- The service should accommodate feedback from domain experts regarding improvements or additions of the metadata kept for the underlying SKGs targeting a better representation of the respective domain (e.g., fixing terminology based on well-established vocabularies or taxonomies). These changes could be part of an initial configuration process for the development of the SKGs that happens offline before the creation of the service instance. Small adaptations of the SKG content should be possible afterwards, as well.

Having determined the aforementioned main requirements, we have then selected an appropriate design approach.

3.2. Design approach

A key objective for the project is to involve researchers closely to the design and development process even from early stages. Within the project, use case representatives are considered to be the domain experts that could be involved in these activities. Hence, we placed significant emphasis on presenting early prototypes to the pilot representatives (following an early and rapid prototyping approach) and gathering feedback related to their specific use cases on many occasions.

In addition, we designed a series of parallel activities to reinforce the previous process. First, we collected early feedback from pilot representatives to understand their use cases and specific needs for the service (documented in Deliverable D1.1). Then, in November 2023, we organised a hands-on workshop in Barcelona to refine these use cases, as a joint exercise between the technical partners and the pilot representatives. We have also drafted an internal document outlining a roadmap for SciLake piloting activities to enhance communication between technical partners and pilot representatives and we formalised the collection of information about domain knowledge spaces and SKG data models using living documents.

The aforementioned approach is part of the design process designed and followed by the project for all software development activities. More details about this process can be found in Deliverable D1.2 ("Initial integrated system").

3.3. Integration into SciLake ecosystem

SciLake's ultimate goal is to provide a suite of customizable components designed to facilitate the creation, interlinking, and maintenance of domain-specific, community-managed SKGs, as well as components to offer a unified method for accessing and querying the included assets. All these components can be deployed and executed across a range of machines, forming a "Scientific Lake", a federated, highly customizable computational infrastructure composed of loosely connected nodes. Each node hosts or executes a subset of the SciLake components tailored to the needs of the respective use case. This Scientific Lake concept aims to enable and support the creation of value-added services that enhance knowledge discovery and other essential research activities.

The smart knowledge discovery service detailed in this report is built upon this foundational concept. Implemented as a value-added service that can be deployed on top of an instance of the Scientific Lake concept, it is also highly customizable by design to meet the specific needs of different scientific domains. This flexibility ensures that the service can adapt to the unique requirements of each domain, thereby providing a powerful tool for researchers to navigate and utilise the vast expanse of scientific knowledge effectively.

The smart knowledge discovery service has two bundles of components: the first comprises a set of tools that can calculate SKG enrichments which can be valuable for the knowledge discovery process; the second provides a powerful user interface (UI) to its end users (i.e., domain experts) offering functionalities that facilitate scientific knowledge discovery. Both types of components are designed to consume the contents of the SKGs that belong to the Scientific Lake mainly via the respective (open) API calls provided by the SciLake nodes. The APIs, among others, support posing Cypher queries to the SKGs exploiting the underlying graph database technology (for details, see D2.1). This offers a lot of expressivity for the type of knowledge that can be retrieved from the SKGs facilitating the implementation of advanced features to serve the domain experts in the context of scientific knowledge discovery.

4. Implementation

In this section, we provide implementation details for all the major components of the service, accompanied by a comprehensive overview of the related development activities. The implementation was guided by the design approach presented in Section 3.2. We describe all technical activities organised by component, based on the components that have been identified during the service design phase, elaborated in Section 3.1. This organisation of the section in component-based subsections was selected to allow for easy navigation and independent reading of the technical details related to each of the components. Additionally, each subsection includes a summary table providing references to the respective code bases, and documentation websites.

4.1. Field extraction component

The field extraction component is responsible for enhancing and refining a suite of classification tools designed to assign Field of Science (FoS)¹ labels to scientific publications. This initiative aims to support the generation of advanced, field-weighted, multi-perspective impact analysis services. Building on a hierarchical classification system initially developed through two European Union studies on Open Science², our approach leverages a publication-based classifier. Such a classification system can be valuable in improving the precision and utility of scientific impact assessments across diverse research fields. In the context of the SciLake project, this classifier and the respective FoS taxonomy will be adapted and expanded to better serve the needs of the SciLake's pilots. Finally, a second classifier responsible for assigning scientific publications to Sustainable Development Goals (SDGs)³ will also be integrated and expanded upon. The SDGs provide a globally recognized framework for addressing critical socio-economic and environmental challenges. By aligning scientific publications with specific SDGs, we can facilitate a more fine-grained and actionable understanding of how research contributes to sustainable development. This view could be of particular importance for some of the SciLake pilots (e.g., the Transportation Research pilot).

4.1.1. FoS Taxonomy

The Field of Science (FoS) Taxonomy is a hierarchical structure of scientific fields that is used as a classification scheme by the FoS Classifier, to automatically classify scientific publications to FoS labels at various levels of detail.

More concretely, the aforementioned FoS classification scheme is underpinned by the OECD disciplines/fields of research and development (FORD) classification scheme, developed in the framework of the Frascati Manual⁴ and used to classify R&D units and resources in broad (first level - L1) and narrower (second level - L2) knowledge domains based primarily on the R&D subject matter. We extend the OECD/FORD scheme by manually linking FoS labels of the

¹ A field of science (FoS) taxonomy in general is a structured classification system that organises scientific disciplines and sub-disciplines. In Section 4.1.1, we define our own structured classification system. In this deliverable we refer to it as FoS taxonomy, unless otherwise specified.

² Concretely, (a) the "Study to support the monitoring and evaluation of the Framework Programme for research and innovation along Key Impact Pathways" under development, and, (b) the MOAP study entitled: "[Monitoring the open access policy of Horizon 2020](#)".

³ The 17 Sustainable Development Goals (SDGs), a call for action by all countries - developed and developing - in a global partnership ([SDG](#))

⁴ <https://www.oecd.org/sti/inno/frascati-manual.htm>

SCIENCEMETRIX⁵ classification scheme to OECD/FORD Level-2 categories, creating a hierarchical 3-layer taxonomy. To facilitate a more fine-grained analysis, we further enhance the Field-of-Science taxonomy with Level-4 (L4), Level-5 (L5) and Level-6 (L6) fields by employing a pipeline of AI techniques.

Generating Level-4 FoS fields: The intuition behind the approach is that venues (conferences or journals) under each Level 3 FoS (e.g., Energy) are creating small communities citing each other. For example, venues that are related to “Renewable Energy” will cite each other more frequently than venues under other Level 3 FoS or venues that are frequent under other subfields of Energy or “general science” venues. We remove from the process “general science” venues (e.g. PlosONE) and create a venue-to-venue citation graph specific to each Level 3 FoS. The nodes on the venue-to-venue graphs are venues and the edges stem from publications that are assigned to the respective Level 3 FoS and are published in venues that exist in the venue-to-venue graphs. We perform community detection to each Level 3 FoS venue-to-venue graph. The resulting communities in each Level 3 FoS graph are the Level 4 FoS fields.

The Level 4 FoS fields are communities of venues and we do not have a label describing each community. The following table provides an example of a Level 4 FoS community of venues.

Level 4 ID	Community of Venues	Manual Annotation
L4_AI_9	acl, naacl, tacl, acm trans asian low resour lang inf process, coling, computational linguistics, emnlp, ijcnlp, int joint conf artif, lang resources evaluation, nat lang eng	Natural Language Processing

Table 1: Example of a Level 4 FoS field under the domain of AI. The Level 4 FoS fields get an ID as seen in the first column. The second column presents a community of venues under a Level 4 FoS.

The third column presents a possible interpretation/label of the community presented in the second column. We cannot manually assign a label to each Level 4 FoS field due to their cardinality and the required domain expertise.

⁵ SCIENCEMETRIX Classification provides a list of Journal Classifications with FoS categories. We get the lower level of this classification to manually link it with the OECD/FORD Level-2 categories.

To provide a label for each Level 4 FoS field, we utilise synthesis publications⁶ and Wikipedia. We identify synthesis publications existing under each Level 4 FoS field (e.g. “*Applying Natural Language Processing and Hierarchical Machine Learning Approaches to Text Difficulty Classification*” is a synthesis publication in Artificial Intelligence). For each synthesis publication we process its title (e.g. lemmatization, stopword removal, POS tagging) and extract Nominal (?) Chunks (NCs)⁷. One aspect of this approach that requires attention, is that the title might also include technologies and very granular fields that would belong to lower levels of our taxonomy. We perform the same text preprocessing in the section titles of the synthesis publications and filter out the common NCs. By performing agglomerative clustering to the resulting NCs and keeping the most representative ones, we end up with NCs that are well-defined and describe the Level 4 FoS field. The following table provides an example of a cluster of NCs in their respective Level 4 FoS field.

Level 4 ID	Cluster of NCs	Wikipedia Assigned name
L4_AI_9	'natural language processing article', 'natural language', 'natural language processing', 'nlp', 'speech language processing', 'computational natural language processing'	Natural language processing/Computational linguistics

Table 2: Example of a cluster of NCs of a Level 4 FoS field. Some clusters have more NCs, however for readability and presentation reasons of this deliverable, we have randomly sampled, if available, 10 NCs. The automatically assigned name to this Level 4 FoS field from Wikipedia is also visible in the third column.

We search each NC from the top-2 most representative clusters to Wikipedia science pages and keep the top-2 most common titles as the name of the Level 4 FoS field.

Generating Level-5 FoS fields: To identify the Level 5 FOS fields, we must first discover the underlying communities formed by the scientific publications within each Level 4 FoS field. Based on the premise that a scientific publication cites thematically related works, we can connect these publications by creating publication-to-publication graphs for each Level 4 FOS

⁶ As synthesis publications, are considered publications with more that 100 references, which usually are literature reviews or surveys, and in general, they try to sum up a scientific field.

⁷ Lemmatization refers to reducing words to their base or root form. Stopword removal refers to eliminating common, non-informative words from text. POS tagging refers to assigning parts of speech to each word in a sentence which indicate the grammatical function of the word within the text and Nominal Chunks (NCs) refer to identifying and extracting noun phrases from text.

field, utilising their citations and references. After creating the graphs, we can again perform community detection to generate communities of publications, which now represent the Level 5 FoS fields. To identify the thematic focus of each community, we apply topic modelling⁸ to the publications within each community. The most frequent topic of each community is considered to be the Level 5 FoS field. The most descriptive n-grams of the rest of the topics under each community are considered to be the Level 6 FoS fields. An example of a Level 5 FoS field (topic) along with Level 6 FoS fields is visible in Figure 1.

By employing open-source Large Language Models (LLMs) and through prompt engineering and few-shot prompting, we can generate a scientific label associated with each Level 5 FoS field (topic). In the following figure a snapshot of the complete taxonomy is presented:

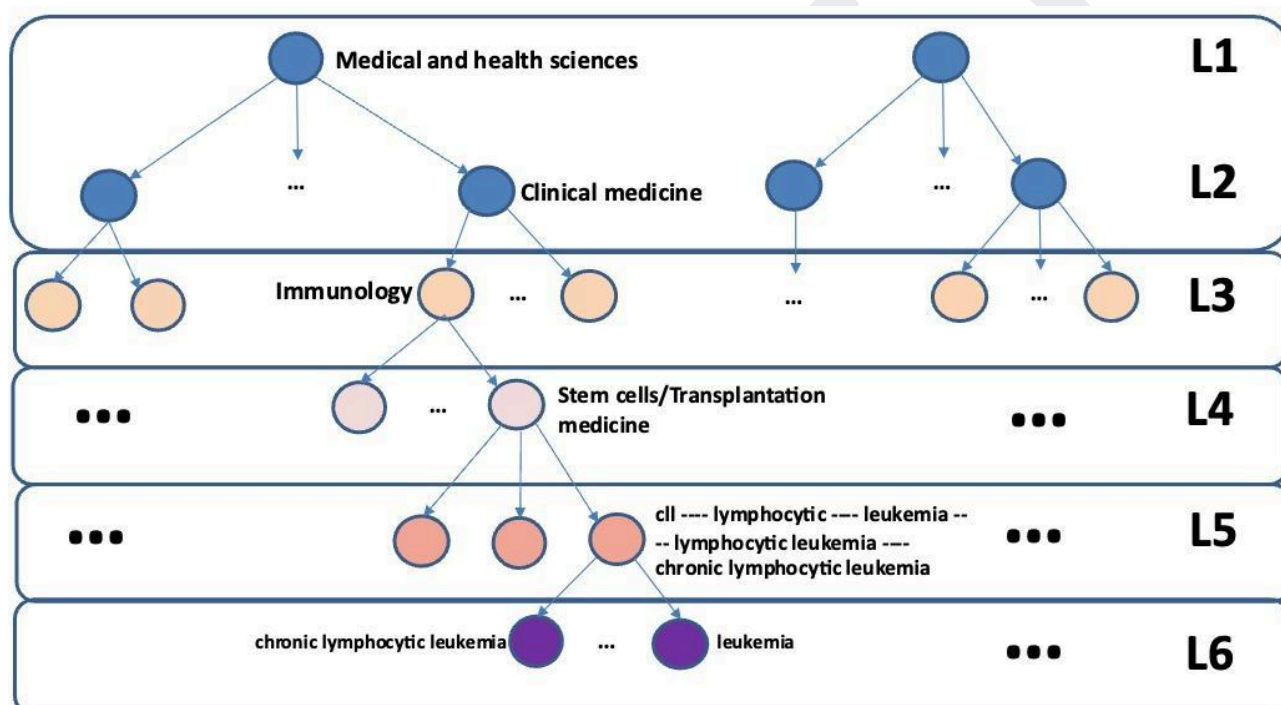


Figure 1: A snapshot of the Field of Science Taxonomy.

More information, results and examples regarding the construction of the FoS taxonomy are included in the relevant papers in Table 3. For the complete FoS taxonomy and LLM prompt refer to the repository link in Table 3. For an overview of Level 1 (L1) - Level 3 (L3) FoS fields refer to [OpenAIRE FoS](#)⁹.

⁸ We utilise BERTopic for the topic model algorithm.

⁹ By typing in the search bar, a user can also browse the L4 FoS fields.

Next steps: The next steps for enhancing the FoS taxonomy are:

- Collaborate with the SciLake pilots to evaluate and refine the automatic assignment of topic labels.
- Create a tool (refer to section 4.1.3) that will be used to identify useful FoS fields for the SciLake pilots.
- Integrate external ontologies/taxonomies of interest, provided from the SciLake pilots, to the FoS taxonomy.
- Evaluate and refine the overall coverage of the FoS taxonomy, by also performing exploratory analysis with other taxonomies and ontologies of science.

The following table summarises the code, Dockerhub, HF space repositories:

Repository: https://github.com/iNoBo/scinobo-fos-taxonomy Paper(s): [3], [4]

Table 3: Links to the repository and publications related to the Field of Science Taxonomy.

4.1.2. FoS Classifier

The FoS Classifier operates on the premise that publications predominantly cite other thematically similar works. It constructs a multilayer network (graph) where nodes represent venues (such as journals or conferences), and edges denote the citation relationships between them. The algorithm classifies a publication P into one or more FoS fields based on the venues of the publications that P references (out-citations) and those that cite P (in-citations). This allows the classifier to effectively classify publications using minimal metadata, relying only on journal or conference names and citation information (i.e. for Level 1(L1) - Level 4 (L4)). Recall from Section 4.1.1 that we have FoS fields for venues.

A snapshot of the multilayer graph of the FoS classifier is presented in Figure 2. In the Figure, the scientific publications (p_i), venues (v_i), and the FoS fields (f_i) are visible and are connected through different types of edges, like *cites* or *cited-by* for venues and scientific publications and *has field* or *subfield* for venues and FoS fields. The classification step consists of propagating information from the venues linked to FoS fields to scientific publications.

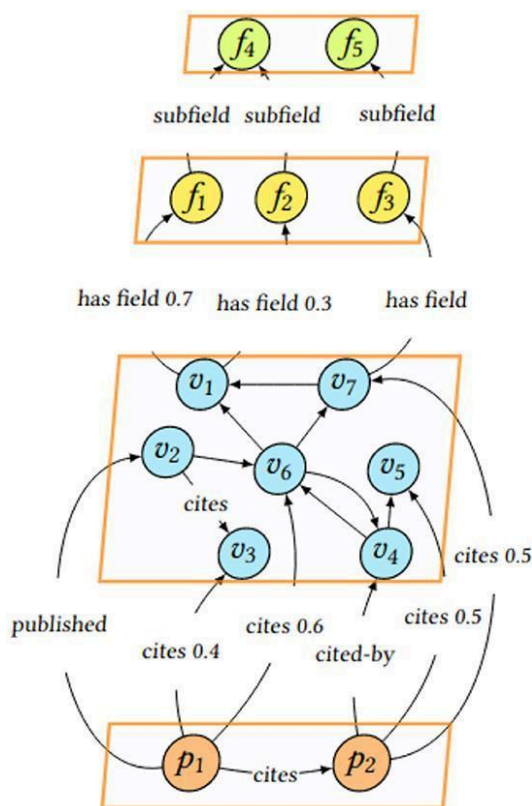


Figure 2: A snapshot of the multilayer graph of the FoS classifier.

There exist multiple ways to back-propagate information from the venue level to the publication level depending on the available metadata, as listed below:

- based on the published venue (namely *Published-by*)
- based on the referenced/cited venues (namely *References*)
- based on the referenced (cited) and citing venues (namely *References+Citations*)

For the readability and demonstration of this deliverable, we omit to provide details for every way to propagate information from the venue level to the publication level. Please refer to the papers listed in Table 4 for further details. The FoS classifier by default utilises an ensemble of all the three aforementioned approaches with additional emphasis to the citations, to infer a publication to Levels 1 up to 4.

Each Level 5 FoS field is associated with a topic and several n-grams as Level 6 FoS fields, stemming from topic modelling as well. To that end, after inferring up to Level 4 and inherently filtering all the candidate Level 5 FoS fields, the inference at Level 5 will adhere to the same principles as those at higher levels, with a key difference: the FoS fields are propagated from the word level to the publication level, rather than from the venue level. We

add the words that co-occur to the multilayer graph of the FoS classifier, drawing edges between them. The weights of these edges are their scores from their respective topic models. We also link the words with their corresponding Level 5 FoS fields in the graph¹⁰. Given a scientific publication, we retrieve its title and abstract. Since the topic modelling algorithm generates topics with words being from unigrams to trigrams, we generate all the unigrams, bigrams, and trigrams in the concatenation of the title and abstract. To classify a publication p , we must map those n-grams to the n-grams in the inference graph¹¹. After the mapping each candidate Level 5 FoS is ranked according to a weighting function¹². The following Figure is provided as a reference point of the abovementioned process.

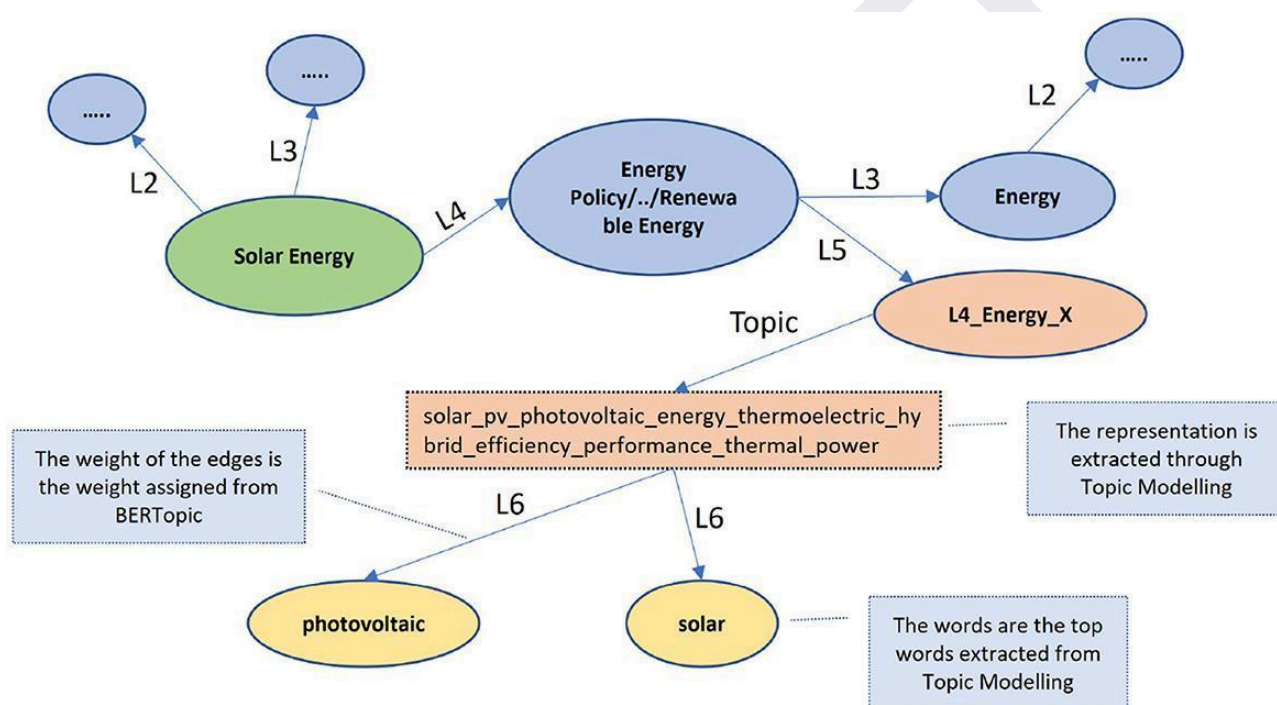


Figure 3: The green node represents a venue, the blue nodes represent FoS fields of all the levels apart from Level 5 and 6. The orange node is a Level 5 (note that it is represented with an ID); and is linked to a topic generated from topic modelling. The yellow nodes are n-grams under that topic.

¹⁰ A complete snapshot of the multilayer graph in JSON format can be downloaded from <https://huggingface.co/datasets/iNoBo/scinobo-fos-inference-graph>.

¹¹ For this mapping we also use semantic retrieval with the inference graph by utilising embeddings, which are numerical representations of words that capture semantic meaning. The embeddings for every n-gram in the FoS inference graph can be downloaded from <https://huggingface.co/datasets/iNoBo/scinobo-fos-graph-embeddings>.

¹² Refer to the listed papers for more information.

Next steps: The next steps at enhancing the FoS classification tool are the following:

- Collaborate with the SciLake pilots to refine and evaluate the classification algorithm by gathering feedback from classified scientific publications.
- Improve the inference speed of the algorithm (e.g. experimenting with binary and scalar quantization of the embeddings of the FoS inference graph).

The following table summarises the code, Dockerhub, HF space repositories:

<p>Code repo: https://github.com/iNoBo/scinobo-fos-classification Docker hub: https://hub.docker.com/r/intelligencenoborders/scinobo-fos-classification Huggingface (HF) space: https://huggingface.co/spaces/iNoBo/scinobo-fos-classification Paper(s): [3], [4]</p>
--

Table 4: Links to code, docker and demo repositories of the Field of Science Classifier.

4.1.3. FoS Taxonomy Mapper

The main objective of T3.1 “Scientific fields extraction from publications” is to provide the functionality required by the SciLake pilots and research communities in general, in order to identify scientific publications of interest. Building upon our FoS taxonomy (section 4.1.1) and FoS classifier (section 4.1.2), we developed an initial version of the FoS taxonomy mapper tool.

The role of the FoS taxonomy mapper is given external Fields of Science (outside of the FoS taxonomy) or indicative keywords/phrases to map them to the most similar or relevant FoS fields of our taxonomy. To achieve the best possible result we will make use of embeddings and more specifically dense retrieval, to retrieve semantically similar FoS fields (instead of using lexical similarity). We employ **Instructor-XL** for our embedding model. More concretely, Instructor-XL is a powerful instruction-finetuned text embedding model that can generate task-specific and domain-tailored text embeddings. It achieves state-of-the-art performance on 70 diverse embedding tasks, making it highly versatile.

Recall that our FoS taxonomy is also a hierarchy. It consists of hierarchical paths starting from Level 1 up to Level 6. To that end, we use Instructor-XL to calculate an embedding per FoS taxonomy path. We exclude from the embedding calculation the first Level, since it is quite general (e.g. natural sciences) and the fifth Level, since we want to directly work with granular ngrams¹³. The rest of the FoS fields are concatenated to create the input to the model. In the following table, examples are presented:

¹³ Keep in mind that this is an initial version of the tool. Through experimentation, finetuning with the SciLake pilots and utilising the Level 5 FoS labels, the embedding path might change.

FoS taxonomy path
clinical medicine/oncology & carcinogenesis/"oncology/infectious causes of cancer"/tumor, malignant, gct, cell tumor, granular cell tumor, cell, granular cell, granular, case, patient
clinical medicine/neurology & neurosurgery/"sleep disorders/sleep physiology"/sleep, osa, obstructive, driver, obstructive sleep, apnea, obstructive sleep apnea, sleep apnea, sdb, sleepiness
"electrical engineering, electronic engineering, information engineering"/energy/fuel cells/urea, oxidation, ni, catalyst, nickel, high, electrochemical, urea oxidation, carbon, fuel

Table 5: Concatenated paths from Level 2 up to Level 6. The ngrams at the lowest level are concatenated with commas.

To calculate embeddings for customised inputs, Instructor-XL offers the following prompt template:

"Represent the [domain] [text_type] for [task_objective].",

where "domain" specifies the text domain (e.g. science, finance, medicine), "text_type" specifies the encoding unit (e.g. sentence, document, paragraph) and "task_objective" represents the objective (e.g. retrieval, classification). We calculate embeddings for every possible path in the FoS taxonomy with the following template:

"Represent the science topic for retrieval:".

Furthermore, we utilise an Elasticsearch¹⁴ cluster where given a query we perform dense retrieval to retrieve the most similar FoS paths from the taxonomy. Given a query, we calculate its embedding using the following template: "Represent the Science sentence for retrieving similar science topics: " and we return the k most similar FoS paths. K is adjustable and can be provided as input to the tool.

In collaboration with the SciLake pilots, we requested each pilot to provide taxonomies and ontologies of interest. As next steps, we are going to use the FoS taxonomy mapper to automatically map the taxonomies of interest to the FoS taxonomy. As a result, we are also going to validate the efficacy of the FoS taxonomy and finetune the taxonomy mapper.

Next steps: The next steps at enhancing the FoS taxonomy mapper tool are the following:

- By utilising the tool, we can evaluate in collaboration with the SciLake pilots, whether the FoS taxonomy offers good coverage to their FoS of interest.

¹⁴ <https://www.elastic.co/>

- Explore different ways of calculating embeddings. For example, we can add to the embedding path the automatically generated topic labels for Level 5 FoS.
- Fine-tune the tool (e.g. each retrieved result is accompanied by a score; finetune this score per external taxonomy.) to improve the retrieval performance in accordance with provided external taxonomies from the SciLake pilots.
- Perform prompt engineering, with the input prompts to the instructor-XL model.

The following table summarises the code, Dockerhub, HF space repositories:

Code repo: https://github.com/iNoBo/scinobo-taxonomy-mapper
Docker hub: https://hub.docker.com/repository/docker/intelligencenoborders/scinobo-taxonomy-mapper/general
Huggingface (HF) space: https://huggingface.co/spaces/iNoBo/scinobo-taxonomy-mapper
Paper(s): -

Table 6: Links to code, docker and demo repositories of the Field of Science Taxonomy Mapper.

4.1.4. SDG Classifier

The General Assembly of the United Nations has adopted a global indicator framework for their Agenda for Sustainable Development. To measure the relevance of scientific articles to the Sustainable Development Goals (SDG) of the UN, an SDG classifier has been developed which classifies a given abstract of a scientific article to one or more SDG categories. A silver corpus of SDG-related articles has been developed using a controlled SDG vocabulary¹⁵ to retrieve metadata of scientific articles found. The controlled vocabulary contains for each SDG category multiple combinations of keyphrases that enable the immediate categorization of an abstract to the corresponding SDG category upon their appearance. By using this controlled vocabulary along with an Elasticsearch installation containing scientific publications, we created a collection of abstracts. Then for each collection of abstracts we applied keyphrase extraction using Textacy's¹⁶ SGRank algorithm to extract additional keyphrases. A human curator reviewed the extracted keyphrases and augmented the controlled SDG vocabulary.

We utilised the above mentioned dataset to train two deep learning models based on a distilled version of BERT [5]. We used a pre-trained version of a distilled BERT model and further fine tuned a Multilayer Perceptron on top of the BERT model using the silver SDG corpus. During training we excluded data from underrepresented categories. We trained two models, one taking into account the vector representation of each token of the input through an attention mechanism and one model that takes into account the vector representation of

¹⁵ <https://zenodo.org/record/4118028#.YvY4GHZBxPb>

¹⁶ <https://textacy.readthedocs.io/en/0.11.0/index.html>

the entire input sequence. As the deep learning models tend to be sensitive to key phrases found in the controlled vocabulary and several SDG categories are underrepresented in the corpus, a guided LDA [14] topic model was also trained using the key phrases found in the vocabulary. During training, a set of key phrases is assigned as a seed to each topic and the topic is labelled with the corresponding SDG category. Guided LDA is thus given a fixed prior probability distribution of word occurrence for each topic. Then LDA is further tuned with the entire silver SDG corpus.

Overall, the classifier consists of two deep learning models and a guided LDA model, combined through a voting strategy to categorise an unseen article abstract. When processing an unseen scientific abstract, if a phrase from the controlled vocabulary is present, the abstract is assigned to the corresponding SDG category. The guided LDA then assigns a score to each topic and its related SDG categories, while the two deep learning models compute a probability distribution across the trained SDG categories. The abstract is ultimately categorised into its SDG classes when the computed scores surpass predefined, adjustable thresholds. In the following Figure, the pipeline of the SDG algorithm is presented:

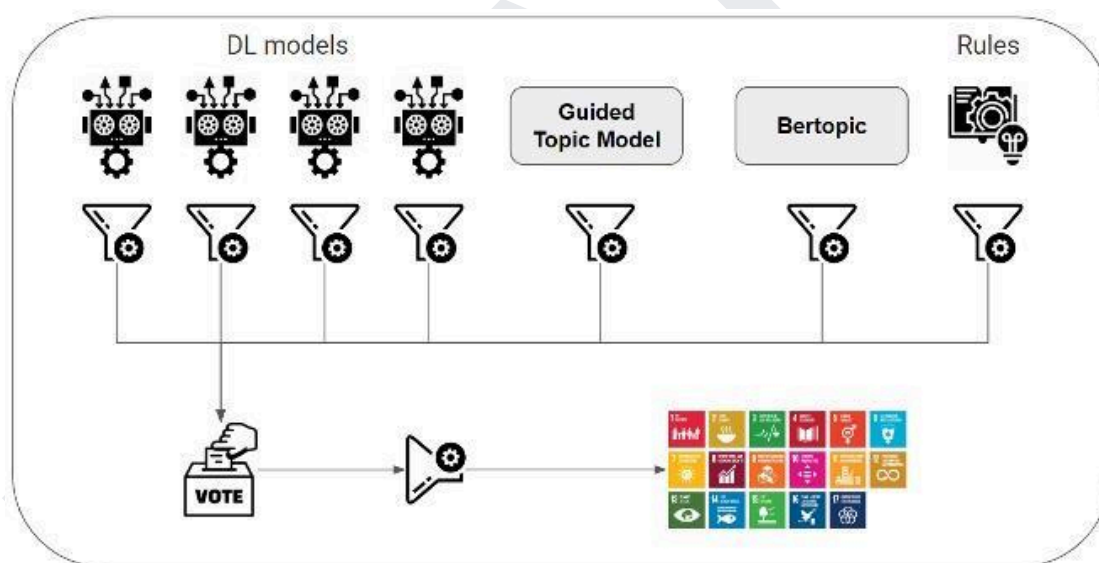


Figure 4: Pipeline of the SDG algorithm.

Next steps: The next steps regarding this activity are the following:

- By utilising the tool, we can evaluate in collaboration with the SciLake pilots the efficacy of the classifier.
- We will use the dataset introduced by [6], to perform a comparative analysis, a thorough evaluation of the tool and the necessary finetuning to improve its performance.

The following table summarises the code, Dockerhub, HF space repositories:

<p>Code repo: https://github.com/iNoBo/scinobo-sdg-classification Docker hub: https://hub.docker.com/repository/docker/intelligencenoborders/scinobo-sdg-classifier/general Huggingface (HF) space: https://huggingface.co/spaces/iNoBo/scinobo-sdg-classification Paper(s): -</p>
--

Table 7: Links to code, docker and demo repositories of the SDG Classifier.

4.2. Citation-based impact analysis component

The citation-based impact analysis component is responsible for calculating indicators of scientific impact of research products by analysing the citation network into which they participate. The calculated indicators are, then, incorporated in the OpenAIRE Graph and, if useful, in some of the domain-specific SKGs. The main use of these indicators is to provide ranking and filtering functionalities for the knowledge discovery service helping researchers identify research products that are key for their investigations. In addition, the indicators are used to provide analytics for the trends related to particular scientific topics. The development activities of this particular component are related to tasks T3.2 “Multi-perspective analysis of scientific impact” and T3.3 “Impact propagation among research objects” of the SciLake project plan.

The following table summarises the code repositories and documentation websites that are related to this SciLake component.

<p>Code repo(s): Citation analysis library (BIP! Ranker): https://github.com/athenarc/Bip-Ranker Publications: [9], [10], [11]</p>
--

Table 8: Links to code repositories and documentation of the citation-based impact analysis component.

Calculating basic indicators. The component currently supports the calculation of an array of citation-based indicators for research products (publications, datasets, software packages, etc.), each capturing a distinct aspect of scientific impact. These indicators, organised by the impact aspect they are capturing, are the following:

- **Influence indicators:** indicators of the "total" impact of each research product, i.e., how established it is in general.

- *Citation Count*: The total number of citations of the research product, one of the most well-known influence indicators.
- *PageRank score*: An influence indicator based on PageRank [7], a popular network analysis method. PageRank estimates the influence of each product based on its centrality in the whole citation network. It alleviates some issues of the Citation Count indicator (e.g., two products with the same number of citations can have significantly different PageRank scores if the aggregated influence of the products citing them is very different - the product receiving citations from more influential products will get a larger score).
- **Popularity indicators**: indicators of the "current" impact of each research product, i.e., how popular the product is currently.
 - *RAM score*: A popularity indicator based on the RAM method [8]. It is essentially a Citation Count where recent citations are considered as more important. This type of "time awareness" alleviates problems of methods like PageRank, which are biased against recently published products (new products need time to receive a number of citations that can be indicative for their impact).
 - *AttRank score*: A popularity indicator based on the AttRank method [9]. AttRank alleviates PageRank's bias against recently published products by incorporating an attention-based mechanism, akin to a time-restricted version of preferential attachment, to explicitly capture a researcher's preference to examine products which received a lot of attention recently.
- **Impulse indicators**: indicators of the initial momentum that the research product received right after its publication.
 - *Incubation Citation Count (3-year CC)*: This impulse indicator is a time-restricted version of the Citation Count, where the time window length is fixed for all products and the time window depends on the publication date of the product, i.e., only citations 3 years after each product's publication are counted.

The previous impact aspects and indicators have been identified and experimentally tested in a series of previous works. More specifically, they have been tested by a large experimental study on ranking publications based on impact indicators [10], a subsequent study that offers improved indicators for the popularity case [9], and another study [11] that resulted in the creation of a dataset that also introduces the concept of impulse indicators.

Moreover, it should be mentioned that all impact indicators are calculated on a deduplicated citation network to avoid counting multiple times citations made by multiple versions of the same product. The nodes of the network are deduplicated using the most recent version of

OpenAIRE's deduplication algorithm for research products¹⁷, hence each node in the network has a distinct OpenAIRE identifier. We also report the scores at PID level (i.e., we assign a score to each of the versions/instances of the product), however these PID-level scores are just the scores of the respective deduplicated nodes propagated accordingly (i.e., all versions of the same deduplicated product receive the same scores). During the previous process, we remove a small number of instances (having a PID) that are assigned (by error) to multiple deduplicated records in the OpenAIRE Graph.

In addition to the values of the indicators, we also offer an "impact class" for each product, which informs the user about the percentile into which the product score belongs compared to the impact scores of the other products in the database. The currently provided impact classes are:

- Class C1 (in top 0.01%)
- Class C2 (in top 0.1%)
- Class C3 (in top 1%)
- Class C4 (in top 10%)
- Class C5 (in bottom 90%)

When it was needed for these classes to be presented in a user interface (UI), for instance in the knowledge discovery portal described in Section 4.4, a combination of an icon and a colour code was used to provide an intuitive way for the end-users to overview the respective information. Moreover, we also provide topic-specific impact classes for the research products. In particular, based on the work done in the Field extraction component (Section 4.1) and the Topic analysis component (Section 4.3), we associate research products with topics and then, for each product and impact indicator, we calculate its class within the respective field (e.g., "in top 1% of the Cancer research field"). Currently, the proof of concept of this functionality is based on concepts provided by OpenAlex, but in the future we will improve the integration with the topics that can be extracted by the Field Extraction component. Currently, for each product, we keep only the three most dominant concepts, based on their confidence scores, and only if the respective scores are greater than 0.3. Then, for each product and impact measure, we compute its class within its respective concepts.

Improving indicators for non-publication products. The assessment of scientific impact traditionally revolves around publications, but in today's research landscape, the impact extends beyond scholarly articles to encompass diverse research outcomes like software and datasets. Although the impact indicators presented in the previous paragraph are calculated for all types of research products, not just publications, the respective scores may not fully

¹⁷ <https://graph.openaire.eu/docs/graph-production-workflow/deduplication/research-products/>

capture the impact aspects of other research products, such as datasets and software. The primary reason is that the impact of a dataset or software is often indicated not only by direct citations but also by other types of mentions. Acknowledging the use of a dataset or software can be done in various ways, and there is no universally adopted set of best practices across all domains and researchers.

Different *forms* of citations and mentions of research artefacts should be considered differently when trying to assess their potential impact or significance. In the context of SciLake, we consider the following forms of citations:

- *Direct citations from publications:* Citations of a research artefact within the reference list of a publication indicate that the artefact has been directly acknowledged and possibly used in the cited work.
- *Direct citations from other research artefacts:* Citations of a research artefact within the content of another research artefact, such as a dataset or software package referenced in a scholarly article.
- *Direct citations to the paper introducing the corresponding artefact:* Citations directly referencing the paper or documentation introducing a research artefact.
- *In-text mentions in publications:* Mentions of a research artefact within the body of a publication, without a formal citation.
- *Mentions in the textual descriptions of other artefacts:* References to a research artefact within the descriptions or documentation of other artefacts, such as software documentation mentioning the use of a specific dataset.

Adapting these multiple forms of citations involves incorporating citations of research artefacts in papers that may occur indirectly, such as in footnotes, to provide a more comprehensive view of their impact.

In addition to the *form* of the citation, other features can be considered for citation indicators to capture a more nuanced understanding of the impact and influence of research artefacts. In particular, as part of this task, we leverage the outputs of the Citation-context assisted replication assessment—see D4.1—to:

- Take into account the type of mention (e.g., usage, mention) of research artefacts, going beyond simply using the section in which they are included as a proxy for the type.
- Explore how to incorporate the predicted polarity and intent of mentions of research artefacts in the computation of their impact. Understanding whether mentions or citations are positive, negative, or neutral, and discerning their intent (e.g., criticism,

endorsement) can provide deeper insights into the perceived impact and reception of research outcomes. For this we will explore the possibility of extending work done for paper citations¹⁸ to the citation of other types of outputs.

In our approach to assessing the impact of research artefacts, we consider two possible methodologies:

- Direct indicators: In this case the research objects are considered independently, assigning different weights to various types of citations or mentions based on their origin or use. This approach draws inspiration from metrics like the u-Index¹⁹, described below, which consider the context and nature of citations to quantify the impact of research artefacts.
- Network metrics: Alternatively, network metrics can be applied to this task, with the different types of relationships between research artefacts encoded as weighted edges in a citation network. In this approach, the entire structure of relationships within the network is taken into account, highlighting the interconnectedness and influence propagation across different artefacts. Here, we will explore how to best integrate the citation information into edge weights in order to apply the existing network metrics calculated by the BIP! service.

The *u-Index* [12] is a resource impact metric based on the number of resource usage citations, the ratio of usage citations to awareness citations, and the age of the resource (in years since first publication) to quantify resource impact. Inspired by author-level impact metrics such as the h-Index, the u-Index distinguishes between citations that indicate "awareness" and those that signify "usage" of informatics resources. Resources with a high ratio of usage citations to awareness citations are likely to be widely used by others and have a high u-Index score.

The formula for calculating the u-Index is as follows:

$$u - \text{Index} = \frac{\text{total \# citations} * \text{usage ratio}}{\text{\# years since first publication}}$$

$$\text{usage ratio} = \text{usage citations} : \text{awareness citations} = \frac{\text{\# usage citations}}{\text{\# awareness citations}}$$

Here, the total number of citations includes any type of citation that a resource has in the citing universe, while the usage ratio represents the ratio of the number of usage citations to

¹⁸ <https://link.springer.com/article/10.1007/s11192-023-04811-5>;
<https://www.researchsquare.com/article/rs-3960194/v1>

¹⁹ <https://www.nature.com/articles/sdata201843>

the number of awareness citations for a given resource. We propose to extend/improve the u-Index by including in the computation of the index different weights based on the different forms of citations, mention types, polarity and citation intent, as described above.

This implementation assumes that the outputs of the WP4 components are already known, so that all forms and types of mentions can be taken into account. However, we will start exploring the weights of the mentions by considering the currently available citation network for research objects available in OpenAIRE. This existing infrastructure will serve as the foundation for expanding the work to encompass additional sources and methodologies for impact assessment beyond publications.

4.3. Topic analysis component

This component provides domain experts with powerful functionalities for monitoring trends in scientific topics. Specifically, it captures and visualises the evolution of topics over time, offering valuable insights. This allows experts to better understand a field of interest, identify emerging areas, and predict which topics are likely to gain attention in the near future. Such capabilities can be useful for researchers and other stakeholders (e.g., officers in funding organisations) for informed decision-making and strategic planning. Furthermore, they provide useful insights for the knowledge discovery process.

The component is based on SciTo [13], a topic monitoring tool that has been developed in the past by ATHENA RC. SciTo relies on a database of mappings between scientific publications and research topics. The contents of this database can be populated by a well-established scholarly data source or using a topic modelling approach on the scientific texts (e.g., titles, abstracts and/or full texts of scientific publications). SciTo's initial version leveraged the results of the execution of the LDA (Latent Dirichlet Allocation) algorithm [14] on the titles and abstracts of a set of scientific publications (more details in the respective publication). Our intention in the context of this project is to give to the end-user the option to select the topics they would like to use for the respective analysis. Therefore, we are experimenting using topics coming from well-established scholarly data sources. More specifically, we have experimented with using concepts from OpenAlex and low-level FoS fields from the FoS Taxonomy being developed by SciNoBo (see also Section 4.1.1).

The main functionality of SciTo is that it offers intuitive visualisations on how scientific topics evolve in the course of time. These visualisations exploit the number of publications related to the topics of interest, their impact in the respective scientific domain, and other information, such as the similarity between different topics (when this is applicable). In the context of the

SciLake project, SciTo functionalities are being adapted, extended and integrated into the BIP! Spaces user interface (UI). This integration aims to create a powerful knowledge discovery tool for domain experts, incorporating all relevant developed features. Currently, the most important topics related to each publication are displayed in the keyword search results, along with a confidence score (when applicable by the classifier used). This score indicates the relevance of the topic to the publication, providing end-users with a clear understanding of the connection's significance. In the UI, this score is represented by a dot, with the coloured portion corresponding to the respective score, ranging from 0 to 1. Furthermore, a summary of the most important topics related to the publications from the end-user's query is displayed at the top of the results, along with the number of related results. The end-user can click on each topic's rounded box to view graphs showing the number of publications related to that topic per year, as well as the corresponding citations. The following figure illustrates a snapshot of the respective user interface.

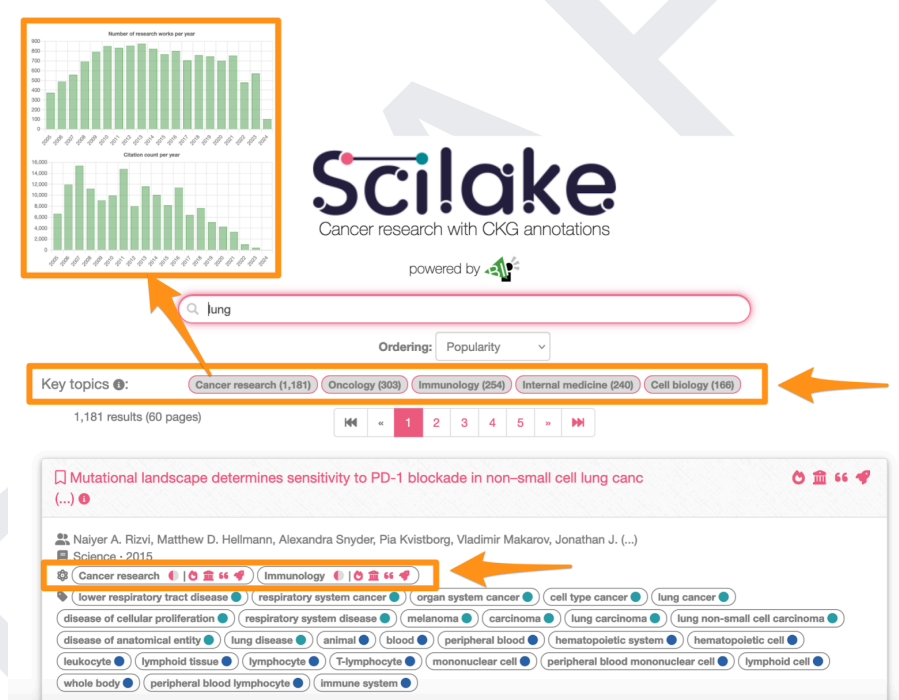


Figure 5: A snapshot of the BIP! Spaces user interface displaying topics.

Additionally, efforts are being made to tailor these functionalities to better serve the specific needs of the SciLake project domains. Pilots have access to the functionalities through early prototype demonstrators, allowing them to experiment with the features and provide feedback on their specific requirements.

The following table summarises the code repositories and documentation websites that are related to this SciLake component.

Code repo(s): https://github.com/athenarc/bip-services Publications: [13]

Table 9: Links to code repositories and documentation of the topic analysis component.

Regarding the next steps, our goal is to further enhance and expand the visualisation functionalities. We also plan to focus on extracting information related to topic evolution by identifying similarities between topics across different time points. We have already started experimenting with Sankey diagram adaptation that can provide useful insights regarding this and we expect to release a relevant prototype demo very soon.

4.4. Knowledge discovery portal

The key objective of this component is to assist researchers and domain experts in the knowledge discovery process. For that reason, it was essential for SciLake to build a user interface (UI) to integrate the previously mentioned components and offer relevant functionalities in this direction. In this section we describe this UI which essentially is a knowledge discovery portal, called "BIP! Spaces".

The following table summarises the code repositories and documentation websites that are related to this SciLake component.

Code repo(s): https://github.com/athenarc/bip-services Publications: [15]

Table 10: Links to code repositories and documentation of the knowledge discovery portal component.

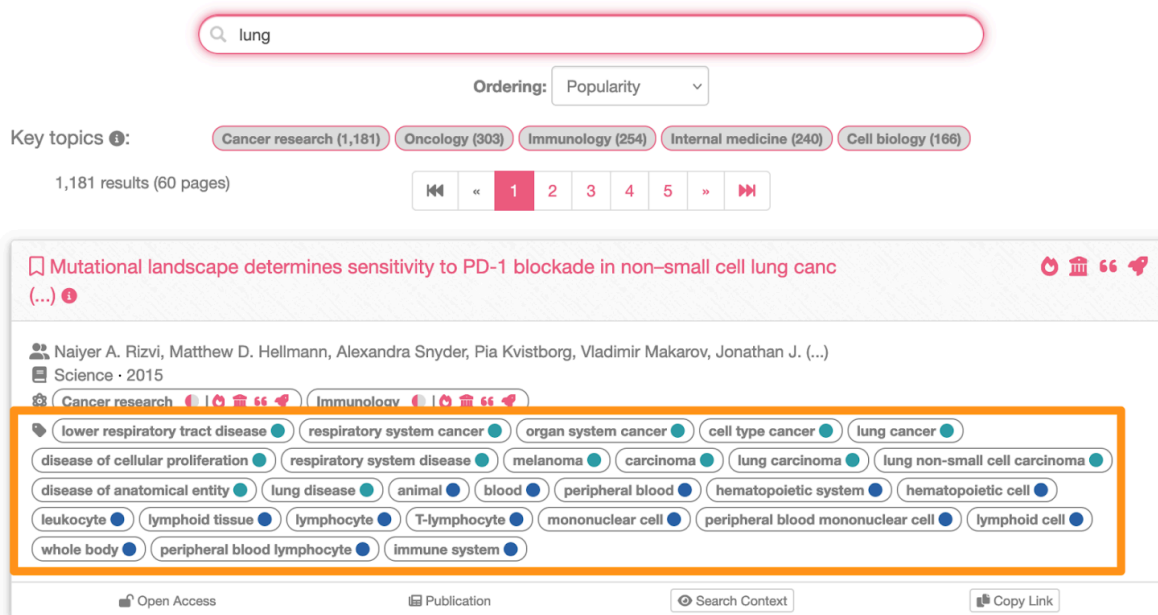
BIP! Spaces is a customizable academic search engine designed so that it can be tailored to meet the specific needs of various scientific domains. Its core functionality includes offering a keyword-based search for research products, such as publications, datasets, software, and other types of research objects. This search can be applied on textual descriptors of these objects (e.g., publication abstracts), or on their contents (in case they are openly available). The retrieved results are also enriched with knowledge included in SKGs, so that the information presented to the end-user will be richer. This assists researchers in identifying valuable resources for their investigations.

The retrieved results can be presented in various ways, such as filtered or organised based on impact indicators, to facilitate the identification of the most valuable items and help prioritise the resources that could be more valuable for a particular investigation of interest. Beyond providing convenient ways to retrieve result lists, the tool also offers various summarization options, including intuitive visualisations. These features enhance the knowledge discovery process and uncover trends or latent relationships between domain-specific entities based on the relevant literature.

Regarding customization capabilities, BIP! Spaces can be tailored to cater to the unique requirements of different scientific fields, ensuring researchers have access to the most relevant and useful tools for their specific areas of study. Specifically, the tool allows the creation of focused "spaces," which are instances of the underlying search engine customised for a particular domain.

Customization options extend beyond basic appearance adjustments, such as using a specific logo, colour scheme, and tagline for the space (just to name a couple indicative options). The entire search experience can be personalised by defining custom pre-selected filters or default ranking methods for search results. The most significant feature, however, is the ability to configure each space to connect with an SKG of interest and consume its contents to offer tailored services to the end-users. This integration allows information for domain-specific entities (and their connection to research products) that is included in the SKG to be offered as annotations of the search results within the UI, enriching the search experience. Furthermore, the SKG contents can further enhance the search experience by providing functionalities like query expansion. For example, synonyms for specific user-provided keywords can be suggested to the end-user during keyword searches, improving the effectiveness and accuracy of the searching mechanism. All customisation options are performed through the admin UI that is responsible for creating and maintaining the respective spaces.

The aforementioned advanced level of customization ensures that BIP! Spaces can adapt to the evolving needs of various scientific domains, offering a highly specialised and efficient research tool. The following figure is offering a snapshot of the basic interface of a space created in BIP! Spaces.



The screenshot displays the Scilake search interface. At the top, a search bar contains the word "lung". Below it, the "Ordering" is set to "Popularity". A "Key topics" section shows categories like "Cancer research (1,181)", "Oncology (303)", "Immunology (254)", "Internal medicine (240)", and "Cell biology (166)". A pagination bar shows "1,181 results (60 pages)" and a set of navigation buttons with "1" highlighted. The main content area shows a search result for a paper titled "Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer". The authors listed are Naiyer A. Rizvi, Matthew D. Hellmann, Alexandra Snyder, Pia Kvistborg, Vladimir Makarov, and Jonathan J. (...). The publication is from "Science" in 2015. Below the title and authors, there are two tabs: "Cancer research" and "Immunology". A large orange-bordered box highlights a set of domain-specific annotations. These include: "lower respiratory tract disease", "respiratory system cancer", "organ system cancer", "cell type cancer", "lung cancer", "disease of cellular proliferation", "respiratory system disease", "melanoma", "carcinoma", "lung carcinoma", "lung non-small cell carcinoma", "disease of anatomical entity", "lung disease", "animal", "blood", "peripheral blood", "hematopoietic system", "hematopoietic cell", "leukocyte", "lymphoid tissue", "lymphocyte", "T-lymphocyte", "mononuclear cell", "peripheral blood mononuclear cell", "lymphoid cell", "whole body", "peripheral blood lymphocyte", and "immune system". At the bottom of the result card, there are buttons for "Open Access", "Publication", "Search Context", and "Copy Link".

Figure 6: A snapshot of the BIP! Spaces user interface displaying domain-specific annotations coming from an SKG.

In the figure, annotations from the SKG connected to the respective space are visible. Specifically, two types of annotations are provided to the end-user: diseases related to each presented research product (displayed in petrol colour) and tissues (displayed in blue colour). By clicking on each annotation's rounded box, the end-user can examine more details about the annotation and its provenance (see also Figure 7). Moreover, in a future version of the UI it will be possible for the researcher to open a detailed page for each annotation, where all research products related to it will be listed together with general information about the annotation. Finally, it should be noted that, currently, the annotations mechanism is also responsible for displaying information related to reproducibility (e.g., datasets or software that is related to each publication), implementing a basic badging mechanism (that will be extended and improved in the future).

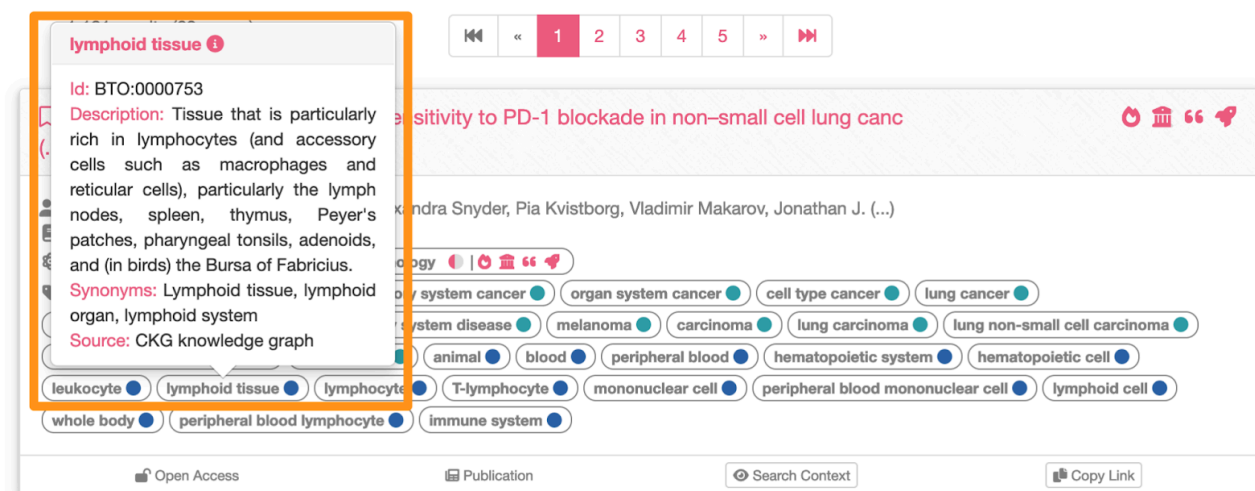
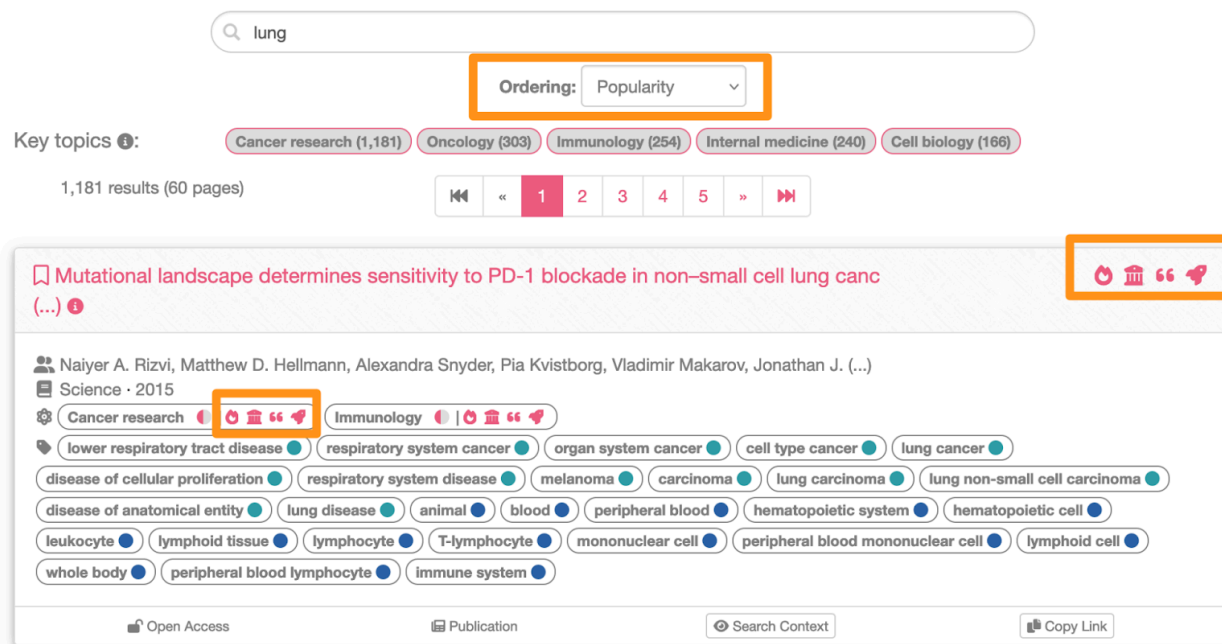


Figure 7: Details for a particular annotation in BIP! Spaces UI.

It's important to highlight how the impact indicators, calculated by the citation-based impact analysis component (Section 4.2), are utilised by the UI. The end-user can choose how search results are ranked using the "Ordering" dropdown list located below the keyword search box. Different ranking criteria can be selected to suit various use cases. For example, ranking by "Influence" or "Citation count" brings fundamental publications to the top, while ranking by "Popularity" highlights more recent results that are currently having a significant impact. Additionally, the end-user can easily assess the percentile rank of each search result by examining the set of impact icons displayed in the top right corner of each result. Moreover, the relative percentile of the respective impact score for each topic associated with the result can be viewed using similar icons within the rounded box of each topic. Figure 8 illustrates these UI elements.



The screenshot displays the SciLake search interface. At the top, a search bar contains the term 'lung'. Below it, an 'Ordering' dropdown menu is set to 'Popularity'. A 'Key topics' section shows filters for 'Cancer research (1,181)', 'Oncology (303)', 'Immunology (254)', 'Internal medicine (240)', and 'Cell biology (166)'. The search results are on page 1 of 60 pages. The first result is titled 'Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer'. The article is by Naiyer A. Rizvi et al., published in Science in 2015. The article has several impact indicators: a red flame icon, a red book icon, a red quote icon, and a red speech bubble icon. Below the article title, there are several topic tags such as 'Cancer research', 'Immunology', 'lower respiratory tract disease', 'respiratory system cancer', 'organ system cancer', 'cell type cancer', 'lung cancer', 'disease of cellular proliferation', 'respiratory system disease', 'melanoma', 'carcinoma', 'lung carcinoma', 'lung non-small cell carcinoma', 'disease of anatomical entity', 'lung disease', 'animal', 'blood', 'peripheral blood', 'hematopoietic system', 'hematopoietic cell', 'leukocyte', 'lymphoid tissue', 'lymphocyte', 'T-lymphocyte', 'mononuclear cell', 'peripheral blood mononuclear cell', 'lymphoid cell', 'whole body', 'peripheral blood lymphocyte', and 'immune system'. At the bottom of the result card, there are buttons for 'Open Access', 'Publication', 'Search Context', and 'Copy Link'.

Figure 8: Impact indicators in the BIP! Spaces UI.

The implementation of these functionalities in BIP! Spaces relies on three main indexes. A Solr²⁰ index supports keyword search, a MariaDB relational database stores domain-independent information from the OpenAIRE Graph, and a graph database instance enables graph queries (using Cypher) on the domain-specific SKG. Communication with the SKG is facilitated through the open API provided by the SciLake ecosystem (see also Section 2.3).

In the upcoming period, BIP! Spaces will be further enhanced to incorporate all the improvements and extensions of the other components discussed in this report (Sections 4.1-4.3). Additionally, several enhancements are anticipated for the current functionalities, such as a more effective query expansion mechanism and improved topic visualisations. Finally, BIP! Spaces will be extended to offer alternative ways to present search results, based on the requirements gathered from the SciLake pilot partners during their experimentation with the respective use cases.

²⁰ Apache Solr: <https://solr.apache.org/>

5. Demonstrated Use Case

In this section, we present a use case to illustrate the value of the knowledge discovery service developed by SciLake, also highlighting the significance of the underlying Scientific Lake infrastructure. These demonstrations are based on the current early beta version of the various components detailed in Section 4.

The following paragraphs describe a use case based on the Cancer research pilot and the current version of the SKG that has been created by the respective experts. It is anticipated that similar demonstrations will be conducted for each of the pilot use cases by the project's conclusion. The BIP! Space²¹ prototype for the Cancer pilot is functional and available for experimentation from the reviewers of SciLake or any other interested individual.

Background. The cancer research pilot focuses on facilitating and empowering the identification of a number of biomarkers with risk-stratifying or predictive impact relevant to providing personalised treatment and care, i.e., precision medicine. The first objective of this pilot is to leverage the functionalities of the Scientific Lake service to create a cancer-specific knowledge graph that can effectively model the respective knowledge space by relying on the defined data sources. The aim is to use several existing knowledge graphs (KGs) from the biomedical field as “building blocks”, and subsequently enrich with cancer-specific knowledge based on the outputs of various SciLake components (e.g., text mining from full-text publications, link prediction).

As a second objective, the pilot will combine this knowledge graph with an adapted version of the functionalities of the scientific merit-driven knowledge space navigation services to facilitate the identification and retrieval of new connecting elements between the biomarkers. The potential applications of a cancer-specific KG are numerous. On the one hand, it will enable researchers to query existing information to deepen their understanding of their findings. On the other hand, it will enable more advanced users to make use of the KG to uncover new information using more advanced algorithms and machine learning approaches. In the context of precision medicine, the cancer specific KG, together with the functionalities integrated on the BIP! Spaces portal created on top of it, is expected to enable new discoveries of patient subtypes and why for example some groups respond better to certain treatments than others. In addition, researchers can use the KG as a benchmark to gauge the novelty of their discoveries in the field.

The current prototype. All figures in Section 4.4 are snapshots taken from the current BIP! Space prototype created for the Cancer research pilot. The appearance of the space has been configured to match the look ‘n feel of the SciLake logo, while various labels (e.g., the main tagline of the space) have been tailored for the cancer research domain (see Figures 6-8). The

²¹ BIP! Space for cancer research: <https://bip.imsi.athenarc.gr/search/cancer-research-ckg>

main discovery parameters (e.g., predefined filters, ordering mechanism) have been set based on an initial understanding of the needs of the specific use case and they are subject to change after experimentation of the domain experts with the UI. The default ordering mechanism is based on the popularity indicator, ensuring that recent results currently receiving a lot of attention are displayed at the top. Moreover, two predetermined filters have been applied in this space: a topic filter that keeps only those research products that are related to the cancer research topic and a filter that keeps only publications in the presented results.

The space is integrated with the current version of the cancer research SKG via the KG engine. This connection primarily enables the enrichment of results with domain-specific annotations, as illustrated in Figures 6-7. Currently, three types of annotations are supported in this space, each derived from a specific Cypher query on the cancer research SKG. Figure 9 demonstrates these three queries within the main administration UI provided by BIP! Spaces. The first type of annotation (presented in dark petrol colour) reveals diseases that are related to scientific publications, the second (in light petrol colour) refers to drugs which are connected to scientific publications, and the third (in blue) provides connections of publications and tissues. All this information is useful for researchers in the field of cancer research and the annotation mechanism of BIP! Spaces gives a quick way to investigate the connection of the existing literature to these domain-specific entities of interest.

In the next period, it is expected that this space will be further extended so that it covers additional types of annotations and features based on the cancer research SKG.

Annotations

Annotation Database

CKG

Annotation Data +

-

Name	Color	Color picker
<input type="text" value="Disease"/>	<input type="text" value="#2c9ca5"/>	<input type="color" value="#2c9ca5"/>
Description		
<input type="text" value="Disease names mentioned in the specific publication."/>		
Query		
<pre>MATCH (d:Disease)-[:MENTIONED_IN_PUBLICATION]->(p:Publication) WHERE p.DOI IN \$dois RETURN (p.DOI), COLLECT({ label: d.name, data: [{ label: "id", value: d.id }, { label: "description", value: apoc.text.replace(d.description, "\\.*?\\", "") }, { label: "synonyms", value: apoc.text.join(d.synonyms, ', ')}]})</pre>		

-

Name	Color	Color picker
<input type="text" value="Drug"/>	<input type="text" value="#9ad7e3"/>	<input type="color" value="#9ad7e3"/>
Description		
<input type="text" value="Drug names mentioned in the specific publication."/>		
Query		
<pre>MATCH (d:Drug)-[:MENTIONED_IN_PUBLICATION]->(p:Publication) WHERE p.DOI IN \$dois RETURN (p.DOI), COLLECT({ label: d.name, data: [{ label: "id", value: d.id }, { label: "description", value: apoc.text.replace(d.description, "\\.*?\\", "") }, { label: "class", value: d.class }, { label: "groups", value: d.groups }, { label: "kingdom", value: d.kingdom }, { label: "subclass", value: d.subclass }, { label: "superclass", value: d.superclass }, { label: "synonyms", value: d.synonyms }]})</pre>		

-

Name	Color	Color picker
<input type="text" value="Tissue"/>	<input type="text" value="#2c61a5"/>	<input type="color" value="#2c61a5"/>
Description		
<input type="text" value="Tissue names mentioned in the specific publication."/>		
Query		
<pre>MATCH (d:Tissue)-[:MENTIONED_IN_PUBLICATION]->(p:Publication) WHERE p.DOI IN \$dois RETURN (p.DOI), COLLECT({ label: d.name, data: [{ label: "id", value: d.id }, { label: "description", value: apoc.text.replace(d.description, "\\.*?\\", "") }, { label: "synonyms", value: apoc.text.join(d.synonyms, ', ')}]})</pre>		

Figure 9: Screenshot from the BIP! Space administration UI showing the current annotations being determined for the cancer research space.

Initial version of the smart impact-driven discovery service

Page 38 of 41

6. Conclusion

In this report, we described the first (beta) release of the smart knowledge discovery service that the SciLake project is developing. This service is built on top of the Scientific Lake that the project is delivering and leverages the contents of the underlying SciLake's Scientific Knowledge Graphs (SKGs). The service is using the SKGs to calculate indicators of scientific impact for research products and offer, based on them, useful functionalities to researchers so that they could prioritise their reading of the relevant literature when searching for a specific subject of interest, uncover latent knowledge that could be instrumental in accelerating their research conclusions, and identify emerging trends related to research topics of interest.

DRAFT

7. References

- [1] UNESCO Science Report: towards 2030. UNESCO Publishing, 2015
- [2] John PA Ioannidis. 2005. Why most published research findings are false. *PLoS medicine* 2, 8 (2005), e124
- [3] Nikolaos Gialitsis, Sotiris Kotitsas, Haris Papageorgiou. SciNoBo: A Hierarchical Multi-Label Classifier of Scientific Publications. *WWW (Companion Volume) 2022*: 800-809
- [4] Sotiris Kotitsas, Dimitris Pappas, Natalia Manola, Haris Papageorgiou. SCINOBO: a novel system classifying scholarly communication in a dynamically constructed hierarchical Field-of-Science taxonomy. *Frontiers Res. Metrics Anal.* 8 (2023)
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT (1) 2019*: 4171-4186
- [6] Dirk U. Wulff, Dominik S. Meier, Rui Mata. Using novel data and ensemble models to improve automated labeling of Sustainable Development Goals. *CoRR abs/2301.11353* (2023)
- [7] R. Motwani L. Page, S. Brin and T. Winograd. 1999. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab
- [8] Rumi Ghosh, Tsung-Ting Kuo, Chun-Nan Hsu, Shou-De Lin, and Kristina Lerman. 2011. Time-Aware Ranking in Dynamic Citation Networks. In *Data Mining Workshops (ICDMW)*. 373-380
- [9] Ilias Kanellos, Thanasis Vergoulis, Dimitris Sacharidis, Theodore Dalamagas, Yannis Vassiliou. Ranking Papers by their Short-Term Scientific Impact. *ICDE 2021*: 1997-2002
- [10] Ilias Kanellos, Thanasis Vergoulis, Dimitris Sacharidis, Theodore Dalamagas, Yannis Vassiliou. Impact-Based Ranking of Scientific Publications: A Survey and Experimental Evaluation. *IEEE Trans. Knowl. Data Eng.* 33(4): 1567-1584 (2021)
- [11] Thanasis Vergoulis, Ilias Kanellos, Claudio Atzori, Andrea Mannocci, Serafeim Chatzopoulos, Sandro La Bruzzo, Natalia Manola, Paolo Manghi. BIP! DB: A Dataset of Impact Measures for Scientific Publications. *WWW (Companion Volume) 2021*: 456-460
- [12] Callahan, A., Winnenburger, R. & Shah, N. U-Index, a dataset and an impact metric for informatics tools and databases. *Sci Data* 5, 180043 (2018).
- [13] Serafeim Chatzopoulos, Panagiotis Deligiannis, Thanasis Vergoulis, Ilias Kanellos, Christos Tryfonopoulos, Theodore Dalamagas. SciTo Trends: Visualising Scientific Topic Trends. *TPDL 2019*: 393-396.

[14] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, null (3/1/2003), 993–1022.

[15] Thanasis Vergoulis, Serafeim Chatzopoulos, Ilias Kanellos, Panagiotis Deligiannis, Christos Tryfonopoulos, Theodore Dalamagas. BIP! Finder: Facilitating Scientific Literature Search by Exploiting Impact-Based Ranking. *CIKM 2019*: 2937-2940

DRAFT