Scientific Lake

DOI 10.5281/zenodo.8403180

Deliverable D1.1: Initial service requirements

Due Date of Deliverable	30/09/2023
Actual Submission Date	30/09/2023
Work Package	WP1
Tasks	T1.1
Type	Report
Approval Status	Submitted
Version	1
Number of Pages	34

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

This document reports on the initial requirements for the services to be delivered by the SciLake project, taking into consideration the results of an initial investigation on the special needs of the SciLake pilot communities (Neuroscience, Cancer research, Transportation, and Energy). It also presents initial ideas on the use cases to be demonstrated by each pilot and possible key performance indicators that can be used to evaluate the effectiveness of the SciLake services in helping researchers in the context of these use cases.

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	1/8/2023	First Draft	Miriam Baglioni, Thanasis Vergoulis
0.2	30/8/2023	Revised version	Miriam Baglioni, Thanasis Vergoulis
0.3	15/9/2023	Peer review comments	Claudio Atzori, Serafeim Chatzopoulos, Mary Melekopoglou
0.4	25/9/2023	Peer review comments addressed	Miriam Baglioni, Thanasis Vergoulis
1.0	30/9/2023	Final version after proof reading	Thanasis Vergoulis

Author List

ORGANISATION	NAME	CONTACT INFORMATION
CNR	Miriam Baglioni	miriam.baglioni@isti.cnr.it
ARC	Thanasis Vergoulis	vergoulis@athenarc.gr

Contributor List

ORGANISATION	NAME	CONTACT INFORMATION
ARC	Serafeim Chatzopoulos	schatz@athenarc.gr
CNR	Claudio Atzori	claudio.atzori@isti.cnr.it
ARC	Mary Melekopoglou	marmel@athenarc.gr

Table of Contents

List of Figures

Abbreviations List

CAD: Connected and Automated Driving

CCAM: Cooperative, Connected and Automated Mobility

DOI: Digital Object Identifier

KG: Knowledge Graph

ORCID: Open Research and Contributor ID

PMID: PubMed IDentifier

RFO: Research Funding Organization

ROR: Research Organization Registry

RPO: Research Performing Organization

SAE: Society of Automotive Engineers

SKG: Scientific Knowledge Graph

SLaaS: Scientific Lake as a Service

V2X: Vehicle-to-everything

1. Executive Summary

SciLake aims to deliver a Scientific Lake infrastructure to facilitate the creation and management of domain-specific Scientific Knowledge Graphs (SKGs) and the development of added-value services on top of them, tailored to cover the special needs of the respective research communities. In the context of the project, four research communities have been selected as pilots to demonstrate and evaluate the aforementioned services: Neuroscience, Cancer research, Transportation research, and Energy research. This deliverable report aims to provide an overview of the activities related to the elicitation and analysis of the initial requirements for the SciLake services from these communities.

First, the methodology followed is discussed: a mixed-methods approach was adopted consisting of online questionnaires and semi-structured interviews. The questionnaires were designed to collect feedback on the current practices, needs, and expectations of the pilot representatives in relevance to scientific knowledge management and discovery and research reproducibility. At least one representative from each of the four pilots participated in this phase. The interviews were conducted to complement and validate the questionnaire feedback. During the interviews, multiple representatives from the four pilots were involved.

Then, the feedback received from the questionnaires is briefly analysed and the main insights gained are discussed. As expected, all pilots have domain-specific data that they would like to utilise in the context of the project. Also, besides the Neuroscience pilot that already has a domain-specific SKG in place, the other pilots aim to build new domain-specific SKGs. In general, most pilots showed interest in almost all the SciLake services, while most of them reported that they do not already have expertise in the underlying technologies, something that indicates that SciLake will bring significant added-value in the respective fields.

After that, an initial use case scenario is presented for each of the pilot cases. These use cases have been created by the SciLake service-providing partners based on the questionnaire responses of the pilot representatives and the feedback that has been collected during the interviews. Quick insights on possible key performance indicators (KPIs) that could be used to track the success of SciLake services in addressing the needs of the pilots in the respective use cases are also presented.

Finally, the report concludes with a summary of the most important results.

2. Introduction

Despite the exponential increase in the output of all scientific fields, the produced knowledge is fragmented and stored in diverse formats. As a result, it is difficult for researchers and other interested stakeholders to utilise this knowledge for advancing science and creating valuable applications for the research community and the society, in general. SciLake's main objective is to assist experts from different scientific fields in organising their domain knowledge in a more structured way, exploiting the advances in Knowledge Graph (KG) and Graph Databases technologies to implement a Scientific Lake (prototype) infrastructure. Moreover, SciLake aims to exploit this Scientific Lake to deliver a set of advanced, added-value services aiming to facilitate activities related to scientific knowledge discovery and research reproducibility. To this end, SciLake aims to deliver three services bundles:

- **Scientific Lake services:** This bundle implements a *"Scientific Lake as a Service" (SLaaS)* infrastructure, which aims to assist users in maintaining, updating, and accessing domain-specific and domain-agnostic scientific knowledge. This knowledge can be well-structured (e.g., in the form of Scientific Knowledge Graphs - SKGs) or, even, completely unstructured (e.g., textual data). The bundle contains a variety of services including (among others): knowledge graph creation services, text mining services, graph mining services, graph querying & analytics services, automatic translation services, scientific content acquisition services, etc.

- **Knowledge discovery services assisted by the use of impact indicators:** This bundle aims to leverage scientific impact indicator technologies to facilitate researchers in navigating the vast knowledge space of the respective scientific domains. The bundle contains a variety of services including (among others): keyword-based search services for research products (publications, datasets, etc), multi-perspective impact-based ranking services for research products, Fields-of-Science (FoS) classification services for research products, topic evolution and trend identification services, and impact propagation technologies. It is worth mentioning that this bundle makes use of the Scientific Lake services (i.e., the first bundle) to access the required scientific knowledge space.
- **Reproducibility assistance services:** This bundle aims to leverage text and graph mining technologies to assist researchers in making their research more reproducible. To achieve this, the bundle offers a variety of services including (among others): services that automatically identify missing links between research products (e.g., publications, datasets, software), services that classify links between research products based on their semantics, services that calculate the reproducibility level of research works (e.g., offering reproducibility badges or indicators) and so on. It is worth mentioning that this bundle makes use of the Scientific Lake services (i.e., the first bundle) to access the required scientific knowledge space.

SciLake pilots are expected to select from the previous bundles those services that seem to be valuable for their own purposes, customise them according to the special needs of their use cases, and combine them in use cases that can demonstrate and evaluate their merit in the determined scenarios.

3. Methodology

In this section, we elaborate on the initial requirement elicitation process that we have followed for the SciLake services. The process was based on two parallel activities:

- The completion (by the pilot representatives) of a questionnaire that was created by the SciLake service-providing partners to get valuable feedback on the pilot needs and collect insights about the desired use cases to be supported by each of the pilots.

- The conduction of a series of interviews to delve into more details regarding the same matters.

As regards the questionnaire, the questions included can be found in Annex 8.1 of this document. The questionnaire was created in Google Forms and was split into six main sections:

1. **Relevant pilot and contact info:** section aiming to gather information about the pilot reported and the respective representative.
2. **Related data:** section aiming to identify data (e.g., KGs, databases, datasets) that could be used to create domain-specific SKGs for the respective pilot community.
3. **SKG creation, interconnection, and federation:** section aiming to collect the respective pilot's needs for specific functionalities related to the Scientific Lake services bundle.
4. **SciLake Smart Services:** section aiming to collect the needs of the respective pilot for smart services/functionalities relevant to SciLake's knowledge discovery and reproducibility assistance services bundles.
5. **Plan of Action:** section aiming to collect any tentative plans of action that the pilot representative has for leveraging SciLake's services.
6. **Outreach:** section aiming to collect contact points or communication channels which are relevant to the respective pilot domain (to facilitate future open consultation activities).

At least one representative for each SciLake pilot (i.e., Neuroscience, Cancer research, Transportation research, and Energy research) was involved in the process and gave feedback to the questionnaires. More specifically, all pilots contributed with one representative, except for the Transportation research pilot that was represented by two people. This was because there are two SciLake partners participating in the Transportation research pilot (ICCS and CERTH) and they decided to give separate feedback to highlight the differences in the sub-domains that they represent. It is worth mentioning that SciLake has also two partners contributing to the Cancer research pilot (CERTH and KI, in particular) but the respective organisations selected to submit a joint feedback.

Regarding the interviews, these were conducted to elicit more detailed and specific requirements from the pilot representatives based on their questionnaire feedback. Therefore, the interviews were semi-structured: the discussion was adapted for each pilot, focusing on clarifying and refining some vague points in the feedback, which were identified from the questionnaires. During these interviews, multiple representatives from the four pilots were involved and multiple interviews have been conducted per pilot.

4. Questionnaires Analysis

This section presents the analysis performed on the responses of the pilot representatives to the questionnaires (see also Section 3). More specifically, each of the following subsections corresponds to one questionnaire section and offers a summary of the participants' responses to the respective questions together with comments on important insights gained.

4.1 Relevant pilot and contact info

This section of the questionnaire aimed to gather general information about the participants and the pilot case they represent. In total, five questionnaires have been gathered for the four pilot cases of the SciLake project. Figure 1 summarises this with a histogram chart.

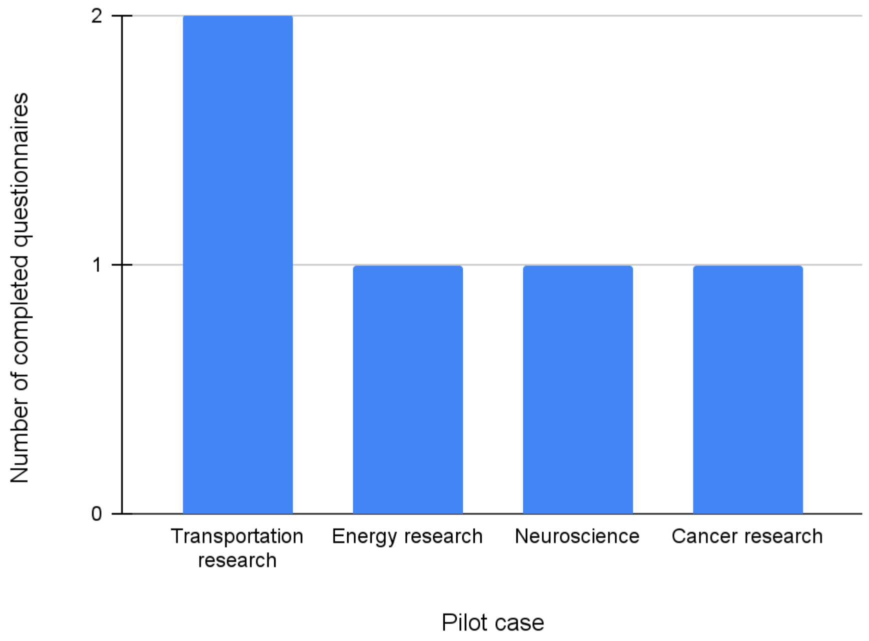

Figure 1. Number of completed questionnaires

The reason why the number of completed questionnaires is different from the number of the pilots is because, for Transportation research, the questionnaire was filled out by two representatives working on the same topic from different perspectives (and different organisations: ICCS and CERTH). For the rest of the report, we will refer to the two answers as "Transportation Research 1" and "Transportation Research 2" (the names are based on the order with which the replies were collected). In contrast, cancer research pilot partners collaborated and filled out the questionnaire in common.

4.2 Related data

This section of the questionnaire investigates the data that the pilot representatives are considering to use in the context of the respective piloting activities. It is worth recalling that, during the pilot activities, each SciLake pilot is expected to leverage data from the OpenAIRE Graph[1] (i.e., the main domain-agnostic Scientific Knowledge Graph included in the Scientific Lake) and from various domain-specific data sources to demonstrate and evaluate the various SciLake services. The domain-specific data sources may already be in the form of Scientific Knowledge Graphs (SKG) or they will be used for the creation of domain-specific SKGs during the project lifetime. Note that, all figures in this section consider responses in the pilot level (i.e., Transportation Research responses are counted only once), therefore we report results given the total number of responses is equal to four (4).

The first question in this section aimed to identify the domain-agnostic, research-related entities which are relevant to the respective pilot. According to the figure, the entities that all pilots commonly reported as relevant were Publications, Datasets, Research software, and Projects. Additionally, Researchers are of interest for three pilots, while RPOs and RFOs are for two. Finally, Venues were not reported as relevant by any of the pilot representatives.

We also inquired about the domain-specific entities that the pilot representatives are considering to use. Three out of five pilot representatives reported such entities; Neuroscience and Transportation Research 2 representatives did not specify any.

[1] OpenAIRE Graph: https://graph.openaire.eu/

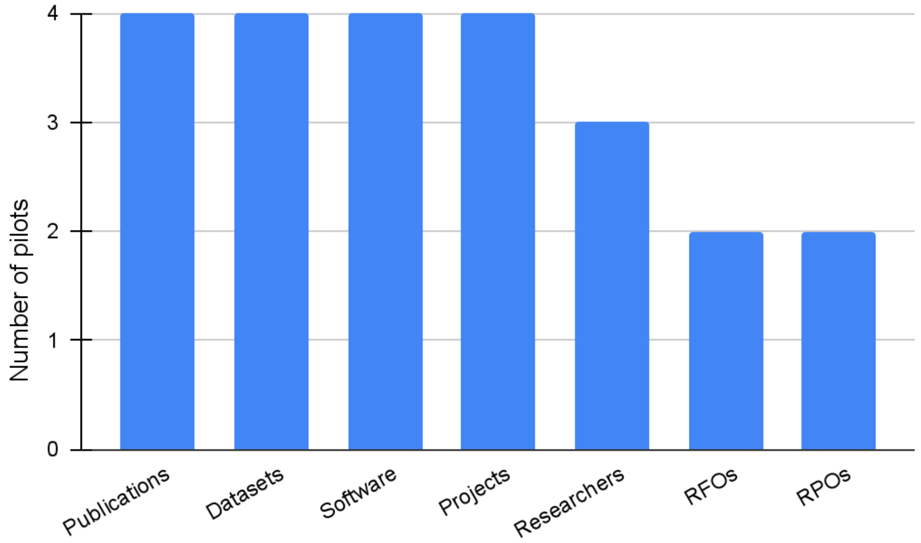

Figure 2. Relevant domain-agnostic entities

According to Transportation Research 1 representatives, the domain-specific entities of potential interest are related to CCAM and cybersecurity; for Energy Research, the domain-specific entities of potential interest relate to renewable energies, climate data, energy supply, energy demand, demographic projections, and climate change; finally, for Cancer Research, domain-specific entities of interest are proteins, genes, pathways, variants, copy number variants, single nucleotide polymorphisms, mutations, clinical impact, survivability, drugs, Cancer types, and phenotypes.

Another aspect considered in this questionnaire section was the existence of domain-specific Knowledge Bases the pilots would like to include in the Scientific Lake. Three pilots out of four reported domain-specific Knowledge Bases of potential interest, as shown in Figure 3. Only representatives of the Energy research pilot did not provide any such resource.

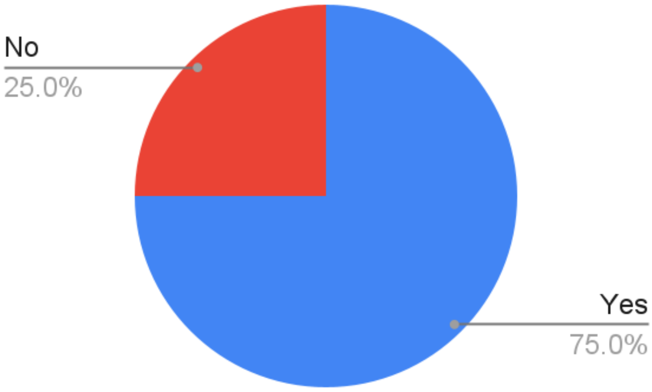

Figure 3. Domain-specific Knowledge Bases

In the following, we briefly describe the outlined domain-specific Knowledge Bases:

- **Knowledge Base on Connected and Automated Driving (CAD)**[2] is the Knowledge Base related to Transportation Research 1. It was initially developed as part of the Horizon 2020 Action ARCADE (Aligning Research & Innovation for Connected and Automated Driving in Europe) and is currently maintained and extended in the frame of the FAME (Framework for coordination of Automated Mobility in Europe) project funded under Horizon Europe, the Knowledge Base gathers the scattered information among a broad network of CAD stakeholders to establish a common baseline of CAD knowledge and provide a platform for a broad exchange of knowledge. It contains information about Projects (>100), twelve Thematic areas, four categories of Regulations, six categories of standards, six categories of evaluation, five categories of data sharing, and policies.
- **TOPOS - Transport Research**[3] is the Knowledge Base related to Transportation Research 2. It aims to showcase the status and progress of open science uptake in transport research. It focuses on promoting territorial and cross-border cooperation and contributing to optimising Open Science in transport research. It features more than 160k research artefacts and is part of the OpenAIRE Graph as a proper subset since it is related to a CONNECT Gateway, thus inheriting all the relationships between TOPOS entities in the OpenAIRE Graph. The persistent identifiers present in the SKG are those present in the OpenAIRE Graph, among which DOI, arXiv, PMID, ROR, ORCID.
- **EBRAINS KG**[4] is a domain-specific scientific knowledge graph that helps researchers find, share, and reuse data in the field of brain research and it was reported by the Neuroscience pilot representatives. It supports rich terminologies, ontologies and

[2] CAD: https://www.connectedautomateddriving.eu/about/

[3] TOPOS: http://beopen.openaire.eu

[4] EBRAINS KG: https://search.kg.ebrains.eu/

controlled vocabularies. It is part of an open research infrastructure that gathers data, tools and computing facilities for brain-related research. Various added-value services have been developed on top of EBRAINS KG.

- **Enrichr KG**[5] is the one relevant for the Cancer research pilot. It is a knowledge graph database and a web-server application that combines selected gene set libraries from Enrichr, a gene set enrichment analysis search engine, for integrative enrichment analysis and visualisation. The enrichment results are presented as subgraphs of nodes and links connecting genes to their enriched terms. It contains information about genes, pathways, phenotypes and much more, and many relations connecting these entities.

The last question of this questionnaire section had the aim to investigate any textual data or general/domain-agnostic resources (i.e. DBpedia) that could be useful in the context of the pilot (i.e. leveraging text mining techniques). Figure 4 shows that only one pilot, i.e., Cancer Research, responded positively to that question. The text corpus, they specified, is the PubMed database that contains more than 35 million citations and abstracts of biomedical literature. The domain-agnostic resources are OpenAIRE, which is already part of the data lake, and ClinVar.

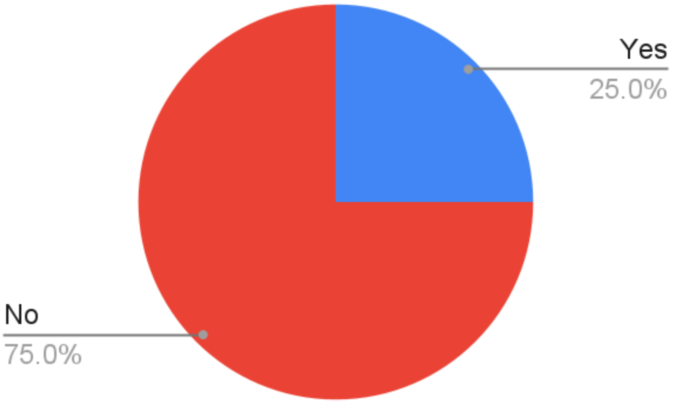

Figure 4. Textual data or domain-agnostic resources

OpenAIRE already collects PubMed metadata and downloads the full text of the open-access publications in the corpus. In the OpenAIRE Graph, each PubMed publication is linked to the bioentities related to it, which are obtained by EMBL-EBIs Protein Data Bank in Europe[6] via its datalinks API[7]. Finally, the set of citations of OpenAIRE will be extended by the POCI set, the

[5] Enrich KG: https://enrichr-kg.dev.maayanlab.cloud/

[6] EMBL-EBIs Protein Data Bank in Europe: https://www.ebi.ac.uk/

[7] Europe PMC APIs: https://europepmc.org/RestfulWebService

OpenCitations Index of PubMed open PMID-to-PMID citations[8] based on the National Institutes of Health Open Citations Collection (NIH-OCC). Its last release features 717,654,703 citations among 29,005,551 bibliographic resources.

ClinVar is a public database that collects and shares information about genetic variations and their effects on human health. It is maintained by the National Institute of Health - National Center for Biotechnology Information (NIH-NCBI). It provides a REST API to query the database programmatically and a full dump updated monthly of its entire dataset downloadable via ftp. It is evident that ClinVar is not a domain-agnostic resource, hence it was reported here by mistake. Clinvar, however, can be kept as an interesting domain-specific knowledge base for the Cancer pilot.

4.3 SKG creation, interconnection, and federation

This section of the questionnaire aimed at collecting the pilot-specific needs for functionalities related to (a) the creation, interconnection, and federation of SKGs and to (b) accessing the Scientific Lake content. In other words, the section aimed to collect feedback for the Scientific Lake services bundle.

The first set of questions was related to understanding if the pilot communities already have domain-specific SKGs they would like to include in the Scientific Lake and whether they are interested in creating new SKGs. The questions also aimed to collect indications about the competencies that the pilot institutions have regarding SKG technologies as well as their needs for computational infrastructure/resources to host and manage the respective SKGs.

According to the responses (see Figure 5), Neuroscience is the only community already owning a mature domain-specific SKG, the EBRAINS KnowledgeGraph[9], to be used for the project. This SKG is already hosted on the premises of the respective pilot partner and the pilot representatives are not interested in creating any other new domain-specific SKG. Furthermore, two pilot communities, Energy Research and Cancer Research, are willing to create new community-specific SKGs and both of them reported that they would require help to accomplish the task. The new SKG could be stored on the pilot partners premises, if the computational resources required are not very demanding. In addition, some of the respective pilot partners are able to ask for resources from alternative infrastructures to host

[8] POCI: http://opencitations.net/index/poci
[9] EBRAINS Knowledge Graph: https://search.kg.ebrains.eu/

the SKGs, in case in-house hosting will not be feasible (e.g., CERTH from the Cancer research pilot considers using resources from ELIXIR-GR's HYPATIA[10] for this purpose).

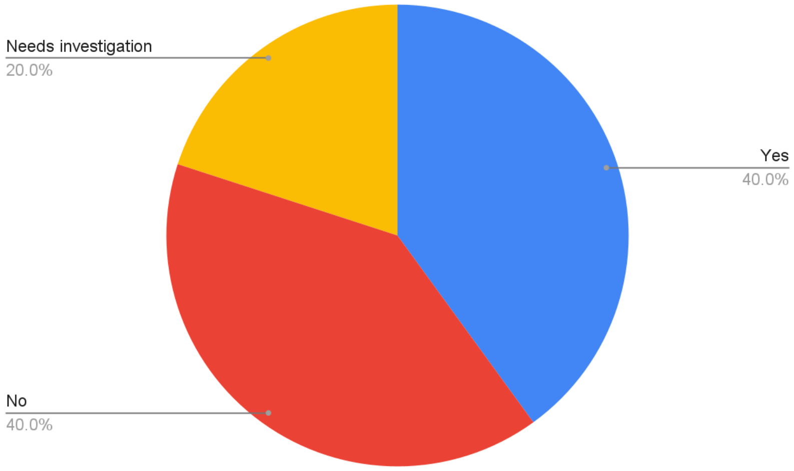

Figure 5. Willing to create new community-specific SKG

Finally, for Transportation Research 1, there is the need to further investigate the issue, however the respective pilot representatives reported that this option would be interesting. Transportation Research 2 reported for the existence of the TOPOS Knowledge Base

The second group of questions in this questionnaire section focuses on tools to facilitate the exploitation of the content of the Scientific Lake.

All the pilots reported that they would like to leverage text mining techniques to extract mentions of domain-specific entities or relationships among them from the scientific texts of the lake.

Figure 6 shows that there is an interest in exploiting tools to create KGs directly from relational databases (on three out of five answers). For Transportation Research 1, the data of interest may include V2X simulation techniques and results, CCAM-related publications, software packages and datasets concerning CAD. For Energy Research, the data of interest are all those related to the energy pilot, such as climate data. Finally, Cancer Research interest lies in PubMed and Clinvar data intending to connect them to Enrichr KG. The Neuroscience community does not have a defined position regarding this functionality.

[10] HYPATIA: https://hypatia.athenarc.gr/

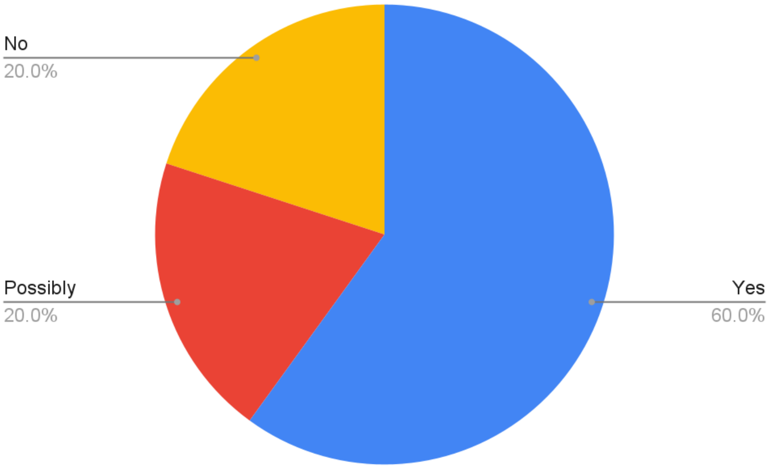

Figure 6. Usage of tools to create Knowledge Graph via direct mapping from the relational database to the graph data model

Figure 7 shows that for one pilot, Energy Research, there is the intention to interlink with other SKGs of the data lake, such as those related to the Transportation pilot. Transportation Research 1 and Cancer Research, are not certain about the need to interlink with other SKGs. If the interlinking will be done for Cancer Research, the pilot would consider the OpenAIRE Graph and possibly the EBRAINS Knowledge Graph. The Neuroscience pilot does not need to exploit interlinking functionalities. Concerning Transportation Research 2, the interest is on enhancing the reproducibility of research in the transport sector, so only domain-specific KGs will be considered.

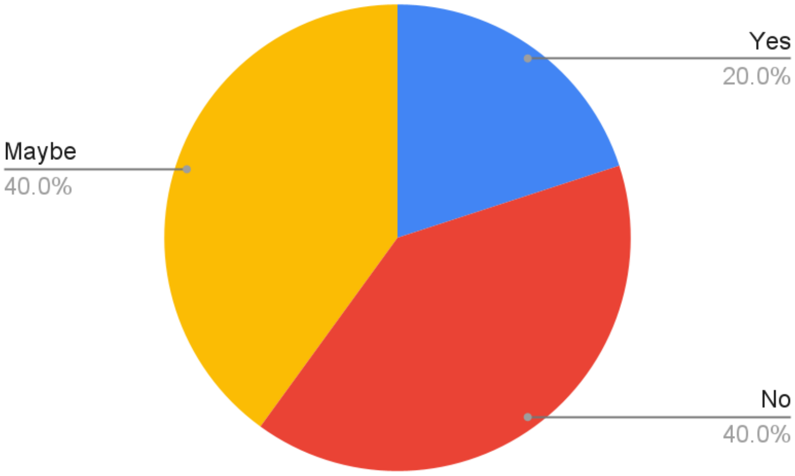

Figure 7. Interest in interlinking with other SKGs in the data lake

As shown in Figure 8, functionalities for searching and browsing in the Scientific Lake are of interest to all the pilots, apart from Neuroscience which did not provide an answer. There is no familiarity with any particular graph query language owned by all the communities. Transportation Research 2 cites Cypher, while from Energy Research, a basic knowledge of SparQL is reported. Transportation Research 1 is willing to allocate resources to using the graph query language since the search and browse functionality is crucial for the pilot.

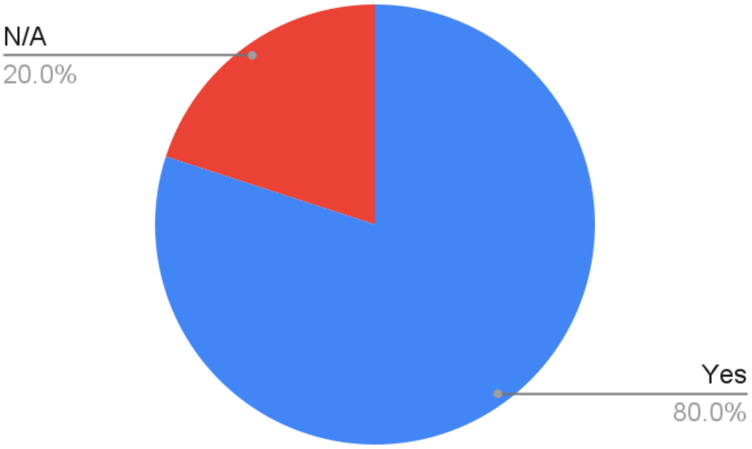

Figure 8. Use functionalities for searching and browsing in the Scientific Lake

Another matter that was reported in the responses was that none of the communities is familiar with or uses third-party tools (i.e. a specific mining module) to perform some of the previous or similar functionalities.

As a last question in this section of the questionnaire, we asked the representatives to provide examples of (high-level) queries they would like to ask the SKG in the context of their pilot.

In particular, Transportation Research (1 and 2 combined) mentioned the following:

- Which CCAM-related projects provide open and reusable V2X simulation code on GitLab?
- How many edge cases can be found in high automation (SAE level 3+) scenarios?
- Queries involving the interconnection of domain-specific entities (i.e. publication, research data etc., related to the topic) with domain-agnostic entities (i.e. projects, researchers, etc.), e.g., all publications of CERTH's researchers in aviation.

For Energy Research, we may expect crossing data queries like which are Europe's top 10% area for energy consumption and population density?

Last but not least, Cancer Research specified queries for interlinking entities or searching for similar data in a different context, e.g., given a biomarker/mutation for cancer:

- does it exist in other cancers?
- how common is it?
- is there a drug against it?
- does it belong to a gene family or function?
- does it interact with other mutated genes in similar cancers?

Finally, we should note that the Neuroscience representative provided no examples of such queries in that question.

We should note that the aforementioned queries seem to be rather specific to the pilots' use cases; however, such queries could be easily answered given the proper modelling of the SKG. It is important to keep in mind that, at this stage of the project, pilots may not be aware of the full potential of SKG systems in evaluating more advanced queries and in running advanced network algorithms. SciLake service-providing partners are planning to inform pilot partners about such opportunities in the context of the continuous co-design process of the project and this will extend the list of potential interesting queries in the future.

4.4 SciLake Smart Services

The respective section of the questionnaire aimed to collect the needs of pilots in relevance to SciLake's impact-based knowledge discovery and reproducibility assistance services bundles. It was split into two subsections, one for each bundle.

i.Impact Driven Services

This subsection investigates the partners' interest in exploiting impact in their pilots and detecting interesting trends in research topics.

All pilot representatives confirmed that they are interested in using impact-driven services in their pilot, but their definition of impact was slightly different in each case. Neuroscience pilot representatives did not provide a domain-specific definition for scientific impact. For Transportation Research (1 and 2), impact is bound to the number of citations, the funding rate, and to the development of beyond-the-state-of-the-art solutions. For Energy Research, it depends on the accessibility of transdisciplinary open data. Finally, for Cancer Research, it depends on the application and practicality of the discovery, especially in a clinical context, on the translational value of the knowledge, and the reproducibility of the methodology.

All the communities would be interested in a knowledge discovery service exploiting multiple research impact indicators. Figure 9 shows the entities of interest for the pilots' representatives.

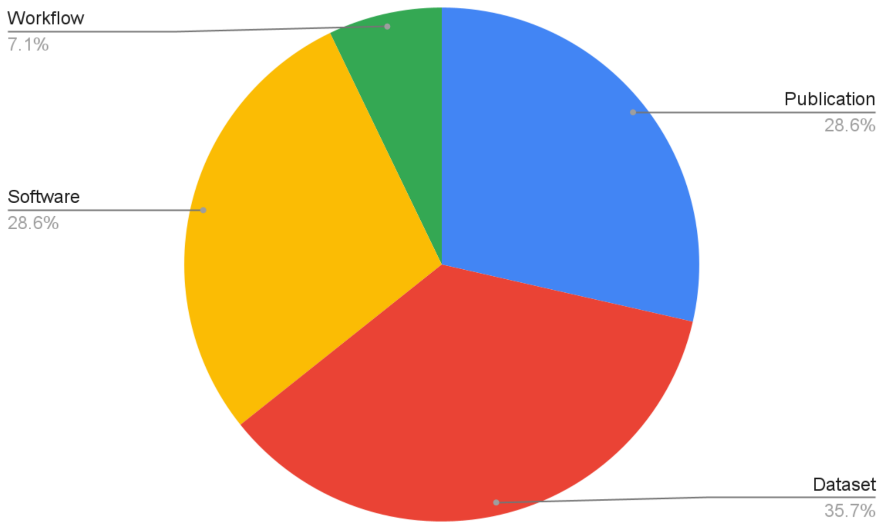

Figure 9. Entities on which to compute the impact indicators

Dataset is the entity type of interest to all the pilots and the only one for Energy Research. Publication and software are of interest to Neuroscience, Transportation Research (1 and 2) and Cancer Research, which is also the only pilot interested in workflows.

All the communities would be interested in propagating the impact scores computed for publication to other, maybe domain-specific, related entities. For Transportation Research 1, the linked entities of interest are dataset and software—no mention of domain-specific entities. For Transportation Research 2, the linked entities of interest are dataset, researcher etc. No mention of domain-specific entities. Energy Research, instead, specified only domain-related entities of interest, such as renewable energies, climate data, and energy demand. The same holds for Cancer Research, where the entities of interest are drugs, cancer types, variants, mutations, and treatments. Finally, Neuroscience pilot representatives did not specify any entity of interest.

In addition, all pilot representatives would be interested in a tool to facilitate tracking and monitoring structural changes in the research topics of their discipline over time (e.g., topic splitting or merging), offering insights on the aggregated impact of each topic and how it evolves, but Neuroscience that did not provide any answer. Energy Research would be interested but unsure how to use such information.

As Figure 10 shows, three out of five representatives are interested in FoS classification for publications. Neuroscience did not provide any answer, while Energy Research is not interested. Other entities other than publication to be tagged with FoS in which Cancer Research would be interested are projects, datasets and software. Transportation Research 1 would be interested in fields like cybersecurity and connectivity, while Transportation Research 2 would be interested in having tagged with FoS all the domain-specific entities.

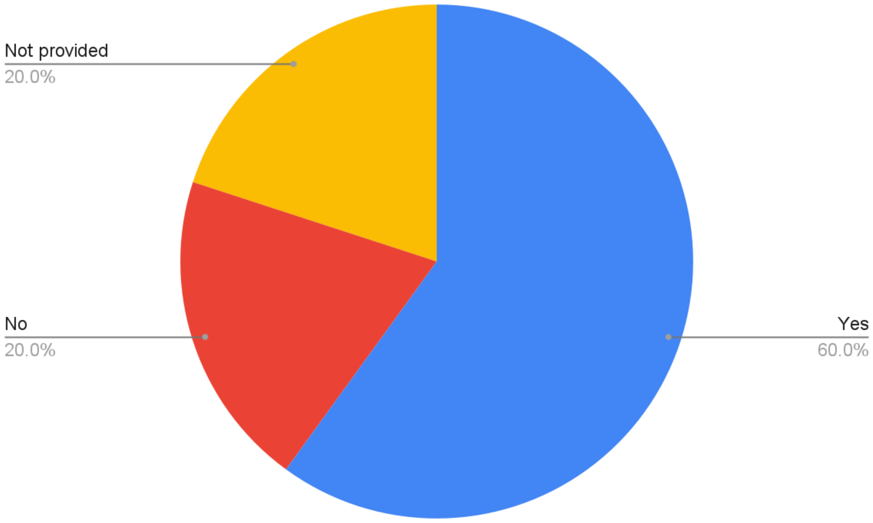

Figure 10. Interest in FoS classification of publications

Two pilot representatives specified domain-specific taxonomies they would be interested in exploiting for subject classification:

2. Transportation Research 1 would be interested in
    a. first level categorisation: Air, Road, Waterborne, Rail, Multimodal
    b. second level categorisation for each entity in the first level: legal/regulatory, technological, transport planning, business modelling, socioeconomics, environmental

3. Cancer Research is interested in the tumour classification specified at the WHO classification of tumours online[11].

Transportation Research 1 and Cancer Research would like to apply this classification to publication, dataset and software and are available to provide feedback.

i. Reproducibility Assistance Services

This subsection investigates the partners' interest in using reproducibility assistance services in their pilot.

All partners are interested in using reproducibility assistance services in their pilots. They would also be interested in a service to offer reproducibility indicators for research works by automatically spotting and isolating mentions of experimental details, configurations, relevant resources (e.g., datasets, software), and other crucial factors for the reproducibility information in the respective publications.

The information essential for assessing the reproducibility of a typical experiment depends on the community. Neuroscience needs further analysis and clarification before answering this question. Transportation Research 1 is interested in spotting the datasets used for experiment simulation and the use cases with multiple parameters applied in the Automated Driving sector. Energy Research identifies reproducibility indicators to be related to the number of geographic areas covered, the spatial data resolution, the aggregation of measures, the covered subjects, time series and transdisciplinary subjects. Transportation Research 2 identifies the crucial factors for reproducibility in datasets, methodologies and algorithms. Finally, Cancer Research identifies software environment, software tools, data format, data availability/deposition, licence, code availability, hardware setup, data documentation, protocols, replicates, reagents (standardised kits vs custom), and training and testing data as the key factors to consider while concentrating on reproducibility matters in their pilot.

All representatives would be interested in a service to recommend links between different research objects (e.g., publications, datasets, software) that are missing from the SKG, since they agree that this will improve the reproducibility of several works in the respective research domain. Figure 11 shows the entities of interest for these new links. Transportation Research 1 specifies links to datasets, software, use cases, simulation data, etc., while Transportation Research 2 focuses more on links between domain-agnostic and

[11] WHO classification of tumours: https://tumourclassification.iarc.who.int/welcome/

domain-specific entities, such as links from researcher to publication or from researcher to Research performing Organizations.

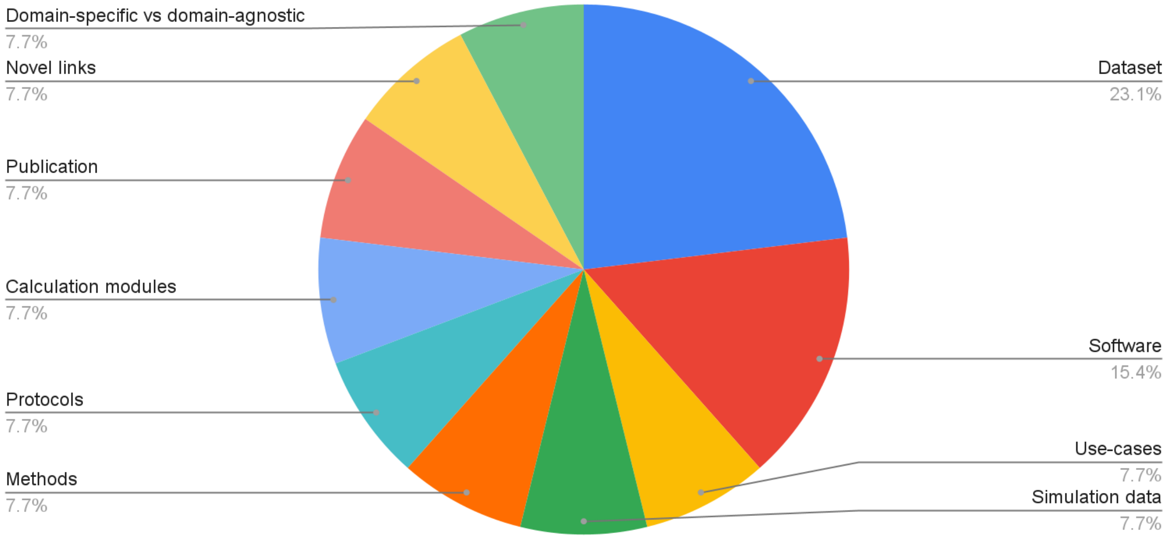

Figure 11. Entities of interest for link recommendation

For Energy Research, the links of interest are within datasets and between datasets and methods/protocols and calculation modules. Finally, for Neuroscience, the links of interest are between datasets, publications and software. Cancer Research would like instead to have novel links. They would be interested in understanding the differences between the novel and the established links but do not specify any entity of interest to find the links.

As Figure 12 shows, three out of five representatives are interested in tools for segmenting scientific papers into their sections/chapters, generalising the variety and variability of section-level titles that can typically be found in the literature. Neuroscience and Transportation Research 2 are not interested. Cancer Research representatives are especially interested if the information is part of the main text: abstract, results, methods, discussion. The supplementary material is of minor importance.

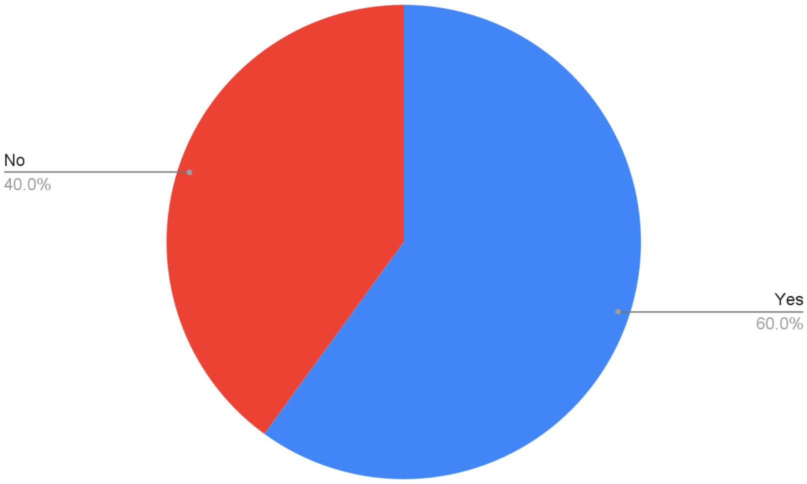

Figure 12. Interest in tools for the segmentation of scientific articles

Moreover, four representatives would be interested in a tool to assess the reproducibility of research based on the analysis of citation statements and the links between different research products (publications, datasets, software). The Energy Research representative is still determining the utility of this functionality. Both representatives for Transportation Research would be interested in using this functionality mostly on English texts, but no other languages are specified. Neuroscience would use it only in English written texts. Cancer Research would also be interested but claims not to have access to such documents for their pilot on SciLake.

Finally, none of the communities uses third-party tools that offer functionalities similar to those described above.

4.5 Plan of Action

The respective section of the questionnaire aimed to collect a tentative action plan from each of the SciLake pilots on how to leverage the SciLake services. Each pilot representative provided a preliminary plan of action that describes which SciLake services are of interest and how they could be used in the context of a piloting activity. It is worth mentioning that, at this point, the aim was not to collect concrete plans but to get valuable initial insights from the pilot representatives. These insights combined with the rest responses in the questionnaire and the feedback received during the interviews (see Section 3) helped SciLake's service-providing partners draft the initial versions of the use case scenarios of the pilots (see

Section 5). Finally, it should be also noted that even these use case plans cannot be considered as final since they will be further refined in the future based on the feedback received in the context of SciLake's continuous service and use case co-design process.

The responses of the pilot representatives in this section indicated that, as it was expected, the SciLake pilots are planning to create or extend their own domain-specific SKGs exploiting the Scientific Lake services, connect them with the OpenAIRE Graph, and then leverage the respective SKG contents together with SciLake services from the impact-based knowledge discovery and reproducibility assistance bundles to facilitate domain-specific use cases. The detailed (initial) plans can be found in Section 5.

4.6 Outreach

The respective section of the questionnaire aimed to collect contact points or communication channels that are relevant to the SciLake pilots. The objective was to collect this information so that it will be possible for SciLake's service-providing partners to interact with a wider audience for each pilot discipline in the context of open consultation activities that may take place in the future.

Neuroscience pilot representatives did not report particular communication channels or contact points; they, instead, suggested leveraging social media like Twitter/X as a communication channel. Cancer Research representatives suggested using the ELIXIR-GR mailing list and the SciLifeLab internal communication as initial communication channels while KI, GMS, and EOSC4Cancer[12] could be considered at a later point. They also suggested the International cancer research community, EOSC4Cancer, Cancer Research Network (KI-Cancer), and StratCan as relevant contact points. Transportation Research 1 representatives reported that they need more time to decide on the right communication channel, while Transportation Research 2 representatives mentioned ECTRI, WEGEMT and EURNEX associations. Finally, Energy Research provided the Enermaps community and the Hotmaps community as related communication channels, while the organisations supporting the OpenAIRE Enermaps gateway[13] as contact points.

[12] EOSC4Cancer project: https://eosc4cancer.eu/

[13] OpenAIRE Enermaps Gateway: https://enermaps.openaire.eu/organizations

5. Use Case Scenarios

This section presents the initial plans for the use case scenarios of the SciLake pilots. The plans were created by the SciLake service-providing partners based on their understanding of the pilot requirements as they were captured by the questionnaire responses and the interviews performed with the pilots during the first months of the project. As such, the plans are subject to refinements during the next period through the continuous co-design process that SciLake envisions to follow. Each subsection focuses on each of the pilots.

5.1 Neuroscience

The Neuroscience pilot is being implemented by the Institute of Basic Medical Sciences of the University of Oslo (UiO) in Norway. This department is heavily involved in the development of the EBRAINS ecosystem[14] which, among others, includes EBRAINS KG[15], a domain-specific SKG for Neuroscience. EBRAINS also offers various front-ends that facilitate metadata curation and scientific knowledge discovery. Based on the above, it is evident that the community behind the Neuroscience pilot is relatively experienced in scientific knowledge management technologies, having various relevant, already implemented services.

As a result, the focus of the Neuroscience pilot will be on activities like:

- ensuring the interoperability of the EBRAINS KG with the OpenAIRE Graph,
- facilitating interlinking of entities between the two graphs,
- enriching EBRAINS KG with contents from the OpenAIRE Graph (e.g., missing connections between datasets and other research-related entities),
- experimenting with SciLake services to implement alternative ways to retrieve content from the EBRAINS KG (e.g., use AvantGraph to support SPARQL or Cypher queries),
- extending EBRAINS services to leverage SciLake impact-based knowledge discovery services and/or reproducibility assistance services to ease everyday routines of the EBRAINS services users (e.g., to assist the curation process with automatic recommendations for metadata enrichments, to provide valuable insights based on impact indicators).

5.2 Cancer Research

[14] EBRAINS: https://www.ebrains.eu/

[15] EBRAINS KG: https://www.ebrains.eu/tools/ebrains-knowledge-graph

The Cancer research pilot is being implemented by the Institute of Applied Biosciences of CERTH in Greece and the Department of Molecular Medicine and Surgery of Karolinska Institutet (KI) in Sweden. The two SciLake partners have extensive expertise in the field of cancer research and are fully aware of existing sources of domain-specific knowledge and the needs of the respective research community. They will work together to:

- create and update a domain-specific SKG by combining information from existing knowledge bases & graphs (e.g., from the OpenAIRE Graph, ClinVAr and Enrichr KG) exploiting, when possible, components from the Scientific Lake service bundle,
- leverage (accordingly customised) SciLake's impact-based knowledge discovery services on top of the aforementioned SKG to facilitate the navigation in the respective knowledge space (e.g., by prioritising reading in the context of clinical interpretation of interesting genomic variants)) and the work of expert curators in further extending (e.g., with additional links) the SKG,
- leverage (accordingly customised) SciLake reproducibility assistance services on top of the respective domain-specific SKG to assist researchers in the field of cancer research in making their research easier to be reproduced.

5.3 Transportation Research

The Transportation research pilot is being implemented by the Hellenic Institute of Transport of CERTH in Greece and the I-SENSE Research Group of the Institute of Communication and Computer Systems (ICCS) in Greece. The two SciLake partners, that have extensive expertise in the field of transportation research, will work together to:

- create a domain-specific SKG integrating technical and scientific reports with datasets, proof of concepts (PoC), and available technical specification documents and guidelines (relying on internal and external dataset sources and the OpenAIRE Graph),
- deploy services to enable smart browsing of the SKG contents based on the impact and reproducibility level of the contained research products aiming to serve types of users (researchers, developers, regulators) having diverse needs,
- deploy services to assist users in identifying the topics of large interest in the Transportation research community and revealing gaps in data and knowledge,
- identify datasets or PoC used rarely but covering specific case study not covered elsewhere,
- exploit text mining and knowledge extraction services that can enable transport researchers to easily comprehend large collections of unstructured text bodies

presenting them in a structured format for identifying meaningful patterns and new insights.

Regarding the source for technical specification documents and guidelines, the pilot representatives reported that EU-based organisations (e.g., ETSI, IETF, CLEPA) should be considered together with US-based (e.g., SAE, UNECE) and Japan-based ones (e.g., jsae, JAMA).

5.4 Energy Research

The Energy research pilot is being implemented by the Institute of Sustainable Energy of HES-SO Valais-Wallis (HEVS) in Switzerland. According to the feedback gathered, the following tasks would be of interest for the Energy research pilot:

- Improvement of the integration/interconnection of existing energy-research-related databases to each other and to domain-agnostic SKGs.
- Deploying services that will assist researchers in prioritising reading of similar case studies exploiting their estimated scientific impact and/or replication level in the Energy research community.
- Deploying services that will classify case studies by geographical location and/or their type.

6. Insights on KPIs

This section summarises insights about possible key performance indicators (KPIs) that can be used to track the effectiveness of the SciLake services in addressing the needs of the use case scenarios identified by the SciLake pilots.

Taking into consideration the initial requirements collected via the questionnaires and the interviews, but also the capabilities of the SciLake services and the objectives of the SciLake project, the following KPIs have been identified to help monitoring the success of the piloting activities (the KPIs are presented together with their initial values at the beginning of the project and their targeted values for the end of the project):

- *K1: Number of SKGs in the lake*
  - Initial: 1 (the OpenAIRE Graph)
  - Target: ≥5

- *K2: Number of SKG nodes created using SciLake services*

    - Initial: 0
    - Target: ≥1mi
- *K3: Number of SKG edges created using SciLake services*
    - Initial: 0
    - Target: ≥10mi
- *K4: Number of textual documents (e.g., full-text manuscripts) in the lake*
    - Initial: 15M (from OpenAIRE data space)
    - Target: ≥30M
- *K5: Number of pilot use case demonstrators for SciLake services*
    - Initial: 0
    - Target: ≥4
- *K6: Number of requests to SciLake services (e.g., via the respective APIs)*
    - Initial: 0
    - Target: ≥10k

These KPIs are expected to help SciLake partners monitor the success of SciLake services in addressing the pilot needs in the context of their use cases. It should be noted that, besides monitoring the previous indicators, we plan to also collect user feedback (e.g., related to satisfaction of the offered functionalities, functionalities usability) and conduct targeted experiments (e.g., to measure performance gains) during the piloting activities, in order to ensure that the services development progresses according to the requirements and expectations of the pilot communities. Finally, it should be also mentioned that the list of KPIs included in this report is subject to slight changes and additions (if needed) based on the expertise that the SciLake partners are going to gain during the project life span.

7. Conclusions

This deliverable report has presented the activities related to the initial SciLake service requirements elicitation and analysis. The work described will drive the research and development of the various SciLake services. Moreover, this report discussed an initial version of the plans for the SciLake pilot use cases and it gave insights related to KPIs that can be used to track the success of the SciLake services in addressing the needs of these use cases.

8. Annexes

8.1 Questions of the questionnaire

1. Relevant pilot and contact info.

This section aims to gather general information about the pilot and the contact person

1. Pilot domain
    a. Neuroscience
    b. Cancer research
    c. Energy research
    d. Transportation research
2. Contact point full name
3. Contact point email
4. Contact point organization

2. Related data.

This section aims to investigate the data relevant to each community that could be interesting to be used for the creation of domain-specific Scientific Knowledge Graphs (SKGs) and the provision of added-value services on top of them.

1. Which of the following (domain-agnostic) entities are relevant to your community?
    a. Publications
    b. Datasets
    c. Research software
    d. Projects
    e. Researchers
    f. Venues
    g. Research performing organizations (RPOs)
    h. Research funding organizations (RFOs)

   i. Other (specify)

2. Please determine domain-specific entities (i.e. chemical compounds, protein complexes, drugs…) that will interest your pilot and can be connected (even indirectly) with domain-agnostic entities of the previous question.

2.1 Domain-specific Knowledge Bases

This part of the related data section aims to investigate if you have domain-specific Knowledge Bases (either owned/maintained by you or well-known ones) that you would like to use in SciLake.

1. Knowledge Base name
2. Knowledge Base URL
3. License
4. Brief description of contents
5. Provide the approximate/precise number of entities per entity type
6. Provide the approximate/precise number of relations, if any
7. Provide any other numerical information you see fit
8. Is the data within the sources identified with persistent identifiers (PIDs)?
   a. Which kind of PIDs?
9. Are the source's data linked to a research product (e.g., a publication, dataset, software) via persistent identifiers?
   a. Which kind of persistent identifiers?
10. If the source is organized as a SKG, provide a description (type of entities and relationships involved)

Do you have other domain-specific KB that you would like to use in SciLake? If yes, determine names and URLs.

2.2 Textual Data and domain agnostic resources

This part of the related data section aims to investigate if you have textual data (e.g., PDF documents) that contain (latent) knowledge that could be useful in the context of your pilot (i.e. leveraging text mining techniques) or a set of general/domain-agnostic resources (i.e. DBpedia) you would like to use

1. Number of documents
2. Languages of the texts
3. License
4. Do you need to extract specific entities from the text (be they provided by the community or already present in the data lake)?
5. Relationships?
6. Do you have a set of general/domain-agnostic resources (i.e. DBpedia) you would like to use? If yes, please provide a list of these resources.

3. SKG creation, interconnection, and federation.

This section aims to collect the pilot needs for specific functionalities related to the creation, interconnection, and federation of SKGs.

1. Do you already own and maintain a domain specific SKG you are willing to share? If yes, please provide a URL describing it
2. Do you need help creating a domain-specific SKG?
3. Do you already have an infrastructure (e.g., a server, a cluster) that could host your domain-specific SKG?
4. Are you interested in using text mining techniques to extract mentions of domain-specific entities or relationships among them from the scientific texts of the data lake?
5. Are you interested in using tools to create knowledge graphs via direct mapping from the relational database to the property graph data model?
    a. On which set of data?
6. Would you be interested in tools to facilitate interlinking between different SKGs in SciLake?
    a. Which SKGs would be of interest in that case?

7. Are you interested in using intelligent and easy-to-use functionalities for searching and browsing the contents of the Scientific Lake?
    a. Are you familiar with a particular graph query language (e.g., SPARQL, Cypher)?
8. Can you provide examples of queries you would like to ask the SKG?
9. Do you already have resources that can be federated with the Scientific Lake?
    a. List of the resources
10. Do you already know or use third-party tools (i.e. a specific text mining module) to perform some of the previous (or similar) functionalities?
    a. Describe each tool for the associated functionality

4. SciLake Smart Services

This section aims to collect the needs of pilots for the smart services/functionalities relevant to SciLake.

4.1 Impact-Driven Services

This part of the Smart Services section is related to the partners interested in computing and exploiting impact in their pilots and detecting interesting trends in research topics.

1. How do you usually consider impact in your community?
2. Would you be interested in a knowledge discovery service exploiting multiple research impact indicators capturing different aspects of scientific impact (i.e. overall influence, current popularity,...) to facilitate searching for domain-specific knowledge? If yes, please select the types of entities.
    a. For which types of research products would you like to have impact indicators?
        i. Publications
        ii. Datasets
        iii. Software
        iv. Other
3. Would you be interested in a tool that would consider the impact scores for publications to propagate then to other (maybe domain-specific) linked entities in your SKG? If yes, which types of entities?

4. Would you be interested in a tool to facilitate tracking and monitoring structural changes in the research topics of your discipline over time (e.g., topic splitting or merging), offering insights on the aggregated impact of each topic and how it evolves?
5. Would you be interested in a functionality to offer subject category classification of publications with Fields of Science (FoS) or other community-related classification schemes and ontologies?
    a. Which entities (e.g. projects) other than publication would you like to be tagged with FoS?
6. Domain-specific taxonomies for subject classification
    a. Are there domain-specific taxonomies/ontologies you would like to exploit for subject classification?
        i. List of resources
        ii. Which entities (e.g., publications, datasets, software) would you like this classification performed on?
        iii. Can you provide feedback for the classification?

4.2 Reproducibility assistance services

This part of the Smart Services section is related to the partners interested in reproducibility assistance services

1. Would you be interested in a service to offer reproducibility indicators for research works by automatically spotting and isolating mentions of experimental details, configurations, relevant resources (e.g., datasets, software), and other crucial factors for the reproducibility information in the respective publications?
    a. If yes, please offer insights about the information that is crucial for the reproducibility of a typical experiment in your domain. You may provide feedback for multiple types of experiments.
2. Would you be interested in a service to recommend links between different research objects (e.g., publications, datasets, software) that are missing from the SKG, as a way to help you improve the reproducibility of your research works?
    a. Which links have the most interest for you?

3. Would you be interested in a tool to facilitate the segmentation of scientific articles into their sections/chapters, generalizing over the variety and variability of section-level titles that can typically be found in the literature?
4. Would you be interested in a tool to facilitate assessing the level of reproducibility of research based on the analysis of citation statements and the links between different research products (publications, datasets, software)?
    a. If yes, would this functionality be useful to also work for non-English texts?
5. Do you already know or use any third-party tools that offer similar functionalities to those described above?
    a. Describe each tool for the associated functionality/Service combination

5. Plan of Action

Plan of action. This section aims to collect a tentative action plan from each pilot for leveraging the SciLake architecture.

1. Please describe a plan of action for your pilot that describes how you could tentatively leverage SciLake services considering your answers to the previous questions. This is just an indicative plan that can be refined later on.

6. Outreach

Outreach. This section aims to collect contact points or communication channels from the pilots to interact with the wider communities in the respective disciplines.

1. Please provide any communication channels from your wider domain that can be used in case we need to get feedback from the wider community.
2. Can you provide any specific contact points interested in this type of interaction?