



Scientific Lake

DOI: 10.5281/zenodo.14748858

Deliverable D5.1: Report on pilot setup and evaluation methodology

Due Date of Deliverable	15th January 2025
Actual Submission Date	16th January 2025
Work Package	5
Type	Report
Approval Status	Submitted
Version	1.0
Number of Pages	54
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 their sole risk and liability.	

Abstract

This report outlines the plans for preparing and implementing the piloting activities of the SciLake project, with a particular focus on the evaluation process designed to assess the technologies developed within the project. The evaluation process aims at evaluating the performance of algorithms and models underpinning core components and services, the value and usability of the respective functionalities in real-world cases inspired by the project pilots, the coverage, accuracy, the utility of the domain-specific SKGs developed for these pilots, and the overall success of the project and piloting activities.



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

Revision history

VERSION	DATE	REASON	REVISED BY
0.0	1/10/2024	First Draft	Andrea Salmi, Jakob Rager
0.1	29/10/2024	Agreement on Structure & References	Andrea Salmi, Jakob Rager, Thanasis Vergoulis
0.2	10/12/2024	Intermediate version	All participants in author list
0.3	15/12/2024	Peer review	Miriam Baglioni, Stefania Amodeo
0.4	10/1/2025	Peer review comments addressed	All participants in author list
1.0	16/1/2025	Final Version after proofreading	Andrea Salmi, Jakob Rager, Thanasis Vergoulis

Author List

ORGANISATION	NAME	CONTACT INFORMATION
HES-SO Valais	Andrea Salmi	andrea.salmi@hevs.ch
HES-SO Valais	Jakob Rager	jakob.rager@hevs.ch
University of Oslo	Ingrid Reiten	ingrid.reiten@medisin.uio.no
University of Oslo	Archana Golla	archana.golla@medisin.uio.no
ATHENA RC	Thanasis Vergoulis	vergoulis@athenarc.gr
ATHENA RC	Sotiris Kotitsas	sotiris.kotitsas@athenarc.gr
ATHENA RC	Sokratis Sofianopoulos	s_sofian@athenarc.gr
ATHENA RC	Serafeim Chatzopoulos	schatz@athenarc.gr
SIRIS	Cesar Parra	cesar.parra@sirisacademic.com
SIRIS	Pablo Accuosto	pablo.accuosto@sirisacademic.com
TUe	Daan de Graaf	d.j.a.d.graaf@tue.nl
KI	Leily Rabbani	leily.rabbani@ki.se
KI	Daniel Hägerstrand	daniel.hagerstrand@ki.se
TUe	Nick Yakovets	N.Yakovets@tue.nl

ICCS	Athanasios Ballis	athanasios.ballis@iccs.gr
CERTH	Afroditi Anagnostopoulou	a.anagnostopoulou@certh.gr
CERTH	Xenofon Kitsios	xkitsios@certh.gr
CERTH	Fotis Psomopoulos	fpsom@certh.gr
CERTH	Georgios Gavriilidis	ggeorav@certh.gr
CERTH	Konstantinos Kardamiliotis	k.kardamiliotis@certh.gr

Reviewer List

ORGANISATION	NAME	CONTACT INFORMATION
OpenAIRE	Stefania Amodeo	stefania.amodeo@openaire.eu
CNR	Miriam Baglioni	miriam.baglioni@isti.cnr.it

Table of Contents

1. Executive Summary	6
2. Introduction	8
3. Piloting Activities Roadmap	9
4. Overview of Pilots	11
Energy	11
Cancer	12
Neuroscience	13
Transportation	13
5. Pilot Preparation & Setup	16
Overview of Preparation & Setup Activities	16
Support Communication Channels	16
Regular Alignment Meetings	16
Knowledge Space Definition	17
Graph Data Modeling	17
Early Prototyping & Testing	17
Targeted Training Activities	18
Report on Related Pilot Partners' Actions	18
Energy	19
Cancer	19
Neuroscience	19
Transportation	20
6. Pilot evaluation	21
Background and General Purpose	21
Evaluation Objectives	21
Methodology	22
Main Evaluation Activities per Component	24
Additional Testing by External Experts	30
Energy	30
Cancer	31
Neuroscience	31
Transportation	32
Evaluation of project activities' success	33
7. Conclusions	35
8. Annexes	36
8.1. Press releases	36

List of Tables

-

List of Figures

-

Abbreviation List

- **API:** Application Programming Interface
- **CCAM:** Cooperative Connected Automated Mobility
- **CLL:** Chronic Lymphocytic Leukemia
- **FoS:** Field of Science
- **KG:** Knowledge Graph
- **KPI:** Key Performance Indicator
- **NLP:** Natural Language Processing
- **RC:** Research Center
- **SKG:** Scientific Knowledge Graph
- **UI:** User Interface

1. Executive Summary

The SciLake project is a collaborative initiative to develop and evaluate domain-specific Scientific Knowledge Graphs (SKGs) and related services to enhance research accessibility and promote innovation across various fields. The project follows an agile, iterative roadmap that includes distinct phases: requirements elicitation, preparation, pilot execution, and evaluation. This approach ensures the project remains aligned with evolving research needs and continuously improves based on stakeholder feedback.

The project comprises four pilot programs focusing on Neuroscience, Transportation, Cancer, and Energy research. Each pilot aims to address critical challenges within its respective domain by creating customized SKGs that integrate scientific knowledge, improve data accessibility, and facilitate knowledge discovery. For instance, the Energy pilot is centred on regional energy planning, promoting sustainable solutions and assisting stakeholders in achieving carbon-neutral energy systems. The Cancer pilot focuses on enhancing personalized medicine for Chronic Lymphocytic Leukemia by identifying biomarkers and improving treatment strategies. In Neuroscience, the project aims to optimize data curation and enhance research by linking curated datasets and publications. The Transportation pilots address Maritime research and Cooperative Connected Automated Mobility (CCAM), developing unified SKGs to improve decision-making and foster innovation in these areas.

During preparation, the project prioritized establishing effective communication and collaboration among the pilot teams and technology developers. Key activities included defining knowledge spaces, identifying essential entities and relationships for each domain, and developing early prototypes of SciLake tools. This process was iterative, with regular feedback from pilot teams driving refinements and improvements in functionalities. Training initiatives, such as webinars and workshops, were also implemented to ensure the project's technological solutions met the specific needs of each domain.

A rigorous evaluation framework underpins the development of SciLake, ensuring that its components and services meet scientific standards and real-world applicability. Evaluation is conducted in three phases: informal feedback gathering, structured testing against benchmarks, and outreach to external experts. The project employs Key Performance Indicators (KPIs) to monitor the success of its activities, including the number of SKGs created and the level of stakeholder engagement. These metrics and community engagement efforts

help validate the effectiveness of SciLake's technologies in real-world research environments and guide future enhancements.

SciLake's integrated approach ensures that its SKGs and services are tailored to specific research needs, adaptable, continuously refined, and aligned with Open Science principles. The project aims to deliver a beta version of its SKGs and associated services by June 2024, marking a significant milestone in its mission to advance scientific knowledge management across diverse research domains.

DRAFT

2. Introduction

The primary goal of scientific research is to deepen our understanding of the world and use this knowledge to improve various aspects of life. However, scientific knowledge is often fragmented and not well-organized to facilitate discovery or the extraction of valuable insights for informed decision-making.

To address these challenges, SciLake introduces a suite of customizable components designed to enable the creation, interlinking, and maintenance of domain-specific, community-managed SKGs. These components provide a unified approach for accessing and querying the contained assets, supporting the development of value-added services to enhance knowledge discovery and streamline processes critical to the broader research community. In addition, to demonstrate the practical application of this approach, SciLake is also developing two discipline-specific, value-added services: one focused on facilitating scientific knowledge discovery, and another aimed at helping researchers improve research reproducibility within specific domains.

To showcase the SciLake services in real-world pilot settings and ensure that the solutions are robust, user-friendly, and effective in addressing the needs of the scientific community, four pilots have been selected, each focusing on a distinct scientific domain: Neuroscience, Cancer, Energy, and Transportation research. The primary objective of these pilots is to demonstrate and evaluate the SciLake customizable knowledge management ecosystem in practical scenarios through tailored pilot demonstrators designed for specific research communities.

The SciLake pilot partners collaborate closely with the component and service providers to test the respective technologies and provide feedback on how they can be improved or adapted to the unique requirements of each domain. Based on this experimentation, they contribute valuable domain-specific insights to support the development and customization of the SciLake ecosystem. Furthermore, the technology-providing partners conduct additional scientific experiments to evaluate various important aspects of their components' performance and effectiveness.

The next sections of this report provide a detailed account of the activities undertaken to prepare for the SciLake piloting and those focused on evaluating the components, services, and pilots and required background information.

3. Piloting Activities Roadmap

In this section we provide a brief overview of the general phases of the SciLake piloting activities. This roadmap includes the pilot preparation and evaluation actions, which are described in more detail in the next paragraphs.

In general, each SciLake pilot was designed to involve activities in the following main phases:

- A. Requirements Elicitation
- B. Pilot Preparation
- C. Pilot Execution
- D. Evaluation

Before delving into each of these phases, it is important to note that SciLake adopts an agile co-design and co-development approach. As a result, the phases are not executed in a strictly linear fashion but are revisited iteratively throughout the project's lifecycle. For example, early prototypes of SciLake components are provided to pilot representatives for testing and feedback during the initial stages, while new requirements may emerge even in the later stages of the project.

Requirements Elicitation. This phase comprised activities focused on collecting domain-specific requirements for the SciLake technologies, leading to the creation of specifications for the design of the respective SciLake components and services. Pilot representatives contributed both general feedback on the respective functionalities and targeted insights tailored to the specific needs of their scientific domain and the identified use cases.

Pilot Preparation. This phase comprises a variety of preparation tasks that can facilitate the pilot execution phase. These include the establishment of support communication channels, the organization of regular alignment meetings between technology-providing and pilot partners, activities that educate technology-providing partners on the domain knowledge of each pilot, and the provision of early prototypes that can be tested to gather initial feedback. Furthermore, various training activities are provided to enhance collaboration and knowledge sharing. During this phase, initial versions of the SKG data models are developed and early versions of the respective graphs for some of the pilots are created.

Pilot Execution. During this phase, the pilot representatives work hand-by-hand with technology providing partners to create and enrich mature versions of the various domain

specific SKGs, execute graph analysis tasks and queries on top of them to reveal hidden knowledge, and experiment with the SciLake value-added services in real-world scenarios.

Evaluation. During this phase, the SciLake partners conduct experiments of various forms (elaborated in next sections) to evaluate the components and services developed within the project, focusing on scientific experiments and on evaluation activities that involve the project pilots and the broader scientific communities they represent.

DRAFT

4. Overview of Pilots

The main objective of WP5 is to showcase and evaluate the SciLake customisable knowledge management ecosystem in real-world scenarios, conducting four pilot demonstrators, each tailored to a specific research community: Neuroscience, Transportation, Cancer, and Energy.

Each pilot overview follows hereunder. The pilot's overview and activities in SciLake are publicly available in different formats through the [project webpage](#) and [press releases](#) developed for project communication.

Energy

The Regional Energy Planning (REP) Pilot, led by HES-SO Valais-Wallis, aims to enhance the accessibility and interconnectedness of scientific knowledge within the energy research domain. Throughout the project activities, this pilot evolved from a broader energy focus to specifically address sustainable energy solutions through regional planning.

The REP Pilot addresses critical challenges in regional energy sustainability, particularly the need for actionable energy plans responsive to different regions' specific energy demands and governance frameworks. A primary focus is on promoting energy sobriety and optimising local and regional energy resources to ensure a secure, resilient, and carbon-neutral energy supply.

The Sustainable Energy and Territory Group of the Institute of Energy and Environment at HES-SO Valais-Wallis, led by Prof. Jakob Rager, manages the pilot. The team adopts a holistic approach that integrates multiple aspects of energy management, including urban planning, social science, marketing, communication, governance, regional development, and engineering. This comprehensive analysis supports effective energy strategies for a wide range of stakeholders, from local administrators to real estate developers, tourism enterprises, banks, and others who aim to incorporate sustainable energy management into their operations.

The scientific knowledge graph (SKG) to be developed as part of this pilot will play a central role in efficiently collecting, organising, and sharing scholarly content. Its use will support semantic modelling, ontology maintenance, and knowledge discovery, which are essential for advancing the field of regional energy planning.

The SKG for energy planning would organise and interconnect data from diverse scientific research, focusing on strategic energy management, resource allocation, policy-making,

sustainability practices, and insights on technical innovations and equipment. This graph will be structured around interconnected nodes representing key topics in energy planning, like energy supply and demand forecasting, renewable energy integration, regulatory frameworks, environmental impacts, socioeconomic factors, and grid resilience. Relationships between these nodes allow users to see how different aspects influence one another, such as the impact of policy changes on energy resource allocation or the interplay between socioeconomic factors and energy demand. By visually and contextually linking research findings, this knowledge graph will support informed, holistic planning and enable quick access to relevant insights for stakeholders involved in sustainable energy development.

Cancer

The SciLake Cancer Pilot has been developed to study how to create accessible, interconnected scientific resources within the cancer research community. Led by teams from the Karolinska Institutet and the Centre for Research and Technology HELLAS (CERTH), this pilot focuses on enhancing the understanding of cancer biology and treatment, specifically targeting Chronic Lymphocytic Leukemia (CLL), the most common form of adult leukaemia.

The Cancer Pilot addresses critical challenges in personalised medicine, particularly identifying key biomarkers for tailored treatment approaches. This pilot aims to build a targeted scientific knowledge graph by combining data from biomedical knowledge graphs with new insights extracted through text and graph mining algorithms. The resulting graph will provide a deeper understanding of CLL's heterogeneous nature and other cancers, ultimately improving treatment strategies and patient outcomes.

The CLL Knowledge Graph (CLL-KG) developed in this pilot is designed to collect, organise, and share data from various sources, including genes, proteins, metabolites, drugs, and clinical trials. This semantic structure of knowledge graphs is anticipated to be pivotal in enabling researchers and clinicians to gain deeper insights into patient subtypes, treatment responses, and emerging therapies.

Building upon the knowledge graph technology, the Cancer Pilot will integrate tools for text mining, entity recognition, and graph mining to enhance knowledge discovery in cancer research and extract meaningful insights for precision medicine.

Neuroscience

The neuroscience pilot, led by Prof. Trygve B. Leergaard (University of Oslo) from the Data & Knowledge Services at EBRAINS (<https://www.ebrains.eu/data/find-data>), aims to connect the European research infrastructure EBRAINS to SciLake's services. The EBRAINS Knowledge Graph (<https://search.kg.ebrains.eu>) contains over 1100 curated datasets and other research outputs, spanning various data generated using a range of structural, functional and molecular methods related to brain research. The collaboration between SciLake and EBRAINS seeks to connect these datasets with a broader pool of neuroscience-related publications available from the OpenAIRE Graph, along with impact metrics and domain-specific metadata automatically calculated and extracted from those publications, to develop an enhanced version of the EBRAINS KG for neuroscience. The new SKG for neuroscience will utilize SciLake data and analytic resources to provide citation counts and other impact metrics that help identify trends in the neuroscience research landscape and optimise data curation efforts in EBRAINS. The integration of EBRAINS datasets in developing the new KG for neuroscience leverages the expertise of the EBRAINS data curation team. Their in-depth knowledge of these datasets is instrumental in validating and improving OpenAIRE functionalities. The SKG developed through the pilot will link research products from several data repositories and publications that will be distributed through the SciLake user interface "Bip! Finder" along with integrated impact indicators.

Transportation

During the project activities, the Transportation Pilot has evolved to focus on two specific areas of interest: Maritime Transportation Research and CCAM Transportation Research. The activities related to each area are designed with distinct objectives outlined in the following paragraphs.

Maritime Transportation Research

The Maritime Transportation Research in SciLake aims to improve the networking of maritime data and scientific knowledge to support safer, more sustainable and more efficient maritime operations. Led by the CERTH/HIT team, this pilot addresses the challenges of fragmented maritime data sources by creating an integrated knowledge graph that unifies diverse maritime information and makes it accessible and usable.

The Maritime Research Pilot focuses on the critical needs of the maritime industry, particularly standardising data, improving safety, and reducing emissions. A key objective is facilitating more effective data-driven decision-making for maritime stakeholders, including operators, policymakers and researchers. This pilot project will help improve vessel tracking, optimise fuel consumption, and support regulatory compliance in different regions by reconciling data sources and incorporating real-time monitoring capabilities.

The SKG being developed for this pilot is critical for compiling, organising and sharing maritime scientific publications and regulations. This knowledge graph plays a central role in supporting the semantic integration of different maritime datasets. It enables advanced analysis and knowledge discovery critical to improving maritime safety and environmental compliance.

Based on a graph structure of *nodes and relationships*, the maritime knowledge graph connects different data types representing entities such as Publications, Regulations, Funding Agencies and Research Projects. The nodes represent individual entities, while relationships illustrate connections between entities, such as Research Field, Citations and Funding. This interconnected structure enables users to gain insights into maritime trends and identify links between publications and researchers, funding sources with publications, and regulations with vessels. By making complex maritime data accessible and interconnected, the Knowledge Graph helps drive the different categories of stakeholders within the maritime transportation research community to gain innovative knowledge discovery capabilities leveraging impact and reproducibility indicators.

CCAM Transportation Research

The Transportation pilot for structured research in the emerging sector of Cooperative Connected Automated Mobility (CCAM) is led by ICCS researchers in Greece. This pilot aims to create an updated scientific knowledge graph specifically for CCAM research, drawing from existing research from co-funded European projects and leveraging metadata from the OpenAIRE Graph while incorporating new developments in the field.

The current contents of the existing subset of the OpenAIRE Graph related to transportation research (<https://beopen.openaire.eu/>) has been rendered obsolete due to its outdated or limited information regarding CCAM. With the sector's rapid evolution, researchers must keep up with new trends, definitions, key terms, and other forms of expert knowledge. To address this gap, detailed taxonomies around CCAM have been identified from other EU-funded initiatives (HORIZON EUROPE projects [SINFONICA](#), [ARCADE](#), [FAME](#)).

SciLake tools leverage such taxonomies and the specific paths identified, with emphasis on critical phrases related to CCAM. The researchers will then be in place to examine the results, fine-tune the chosen paths from the taxonomy, and identify new expert knowledge. As a result, new datasets are expected to be unveiled, significant papers to be located, and current research trends to emerge.

In parallel, an existing knowledge base from the EU-funded project FAME (<https://www.connectedautomateddriving.eu>) provides valuable data that can be incorporated into the project's developed models. This repository contains raw, unstructured information covering a wide range of aspects around CCAM, from critical enabling technologies to legal documentation and general thematic categories. However, it needs to be appropriately processed to be ready for use by the project's models. The ultimate goal is to fine-tune the targeted extracted results to assist researchers in their work on CCAM.

DRAFT

5. Pilot Preparation & Setup

The execution of the pilots is of great importance for validating and improving the SciLake services. Thus, preparation activities were necessary to ensure the successful execution, support, and evaluation of the SciLake pilots. In practice, various factors (e.g., familiarity with relevant technologies or software, unknown requirements) can influence the implementation of the planned pilots and impact the respective activities. In the following sections, we first elaborate on the preparation and setup process we followed to anticipate and mitigate any possible setbacks and problems. Then, we report on related actions performed by the pilot partners.

Overview of Preparation & Setup Activities

This section provides an overview of the most essential activities to prepare and set up the SciLake pilots.

Support Communication Channels

Since the beginning of the project, a series of communication channels were established to support piloting activities. These channels included designated contact points from each of the SciLake services or components implementation teams, mailing lists (e.g., WP-specific lists), and dedicated Slack channels for discussing relevant matters. Leveraging them became a key method for addressing requests for explanations and clarifications related to the piloting activities.

Regular Alignment Meetings

The project established regular virtual meetings to align the implementation and piloting activities. Two central meetings were held: bi-weekly WP5 meetings focused on reporting and discussing the progress of piloting activities and monthly joint technical meetings centred on implementation efforts. Both types of meetings included participants from both groups of partners, facilitating interactions and collaboration. Consequently, these regular meetings also served as a discussion forum, offering opportunities to provide clarifications and explanations whenever needed.

Knowledge Space Definition

To facilitate the design, implementation, and appropriate configuration of all SciLake components, each pilot team prepared a concrete and detailed definition of their respective domain's Knowledge Space, including descriptions of the domain, key input sources (such as existing knowledge graphs, databases, and text corpora), domain-specific vocabularies, taxonomies, and the languages used to describe the knowledge. In addition, in the same activity, pilot representatives created and customised pilot-related OpenAIRE Gateways to help them identify domain-specific research products (like publications, datasets, and software). All this information helps technology partners to become familiar with the scientific domains of the pilots, facilitates the process of drafting the respective SKG data models, and enables the identification of data sources and approaches for the population and enrichment of the SKG contents.

Graph Data Modeling

Continuing the work mentioned in the previous paragraph, each pilot team identified key domain entities, the relationships between them, and the metadata that would be valuable to be associated with each entity and relationship. This information defined the first version of the Graph Data Model that should be used for the domain-specific Scientific Knowledge Graph (SKG), which the respective pilot will need to develop with the help of the SciLake component and service providers. The model identified knowledge sources, and OpenAIRE Gateways form the foundation for building the SKGs.

Early Prototyping & Testing

The technology-providing partners prepared and made available early prototypes of the services and components for the pilot representatives so that they could test the developed features and functionalities. Pilot representatives tested those prototypes, and this early engagement allowed them to gain expertise and become familiar with the underlying technologies. Additionally, these efforts provided valuable feedback, contributing to the further development, enhancement, and customisation of these components and services.

In the context of those activities, pilot representatives interacted with scripts, Web-based UIs, and APIs. Furthermore, they provided sample collections of research publications for text segmentation, automatic translation, field extraction, and named entity recognition purposes.

Targeted Training Activities

SciLake partners have organised several training activities to enhance collaboration and knowledge sharing. First, a series of *internal webinars* was conducted to strengthen the connection between the research communities involved in the pilots and the technical partners working on the SciLake infrastructure. These webinars also aimed to better inform the pilot representatives about the underlying technologies. A survey was conducted to identify the most interesting topics and identify seminar subjects involving the pilot representatives. SciLake's technology-providing partners invested effort in preparing the materials required for these training activities. Reports of these meetings are published in the [Training blog](#) section of the SciLake website:

- <https://scilake.eu/amplifying-valuable-research>
- <https://scilake.eu/openaire-graph-scilake>
- <https://scilake.eu/avantgraph>
- <https://scilake.eu/science-no-borders>
- <https://scilake.eu/machine-translation>

Furthermore, surveys were conducted after each webinar to gather feedback and guide future development across work packages.

During the plenary meeting in Barcelona in November 2023, pilot and technology-providing partners participated in a *matchmaking and collaboration workshop* designed to map SciLake services to pilot use case needs. This process created a structured map outlining interactions and responsibilities across work packages. The workshop included a value proposition canvas activity that helped align services with pilots' needs. The exercise mapped three key elements: the challenges each pilot faced ("pains"), what they hoped to achieve ("gains"), and their core tasks ("jobs"). This analysis ensured that SciLake's ecosystem would provide meaningful value to each pilot's requirements.

Report on Related Pilot Partners' Actions

This section provides an overview of the SciLake pilot partners' actions related to pilot preparation and setup. It is worth mentioning that HES-SO, as the WP5 leader, was responsible for coordinating the respective activities, managing the monthly WP5 meetings, and participating in the monthly WP leaders' meetings. HES-SO organised activities such as creating the service-to-pilot mapping and the value proposition canvas during the plenary meeting in Barcelona in November 2023 to facilitate collaboration. HES-SO also coordinated the reporting of WP5, supported pilot customisation, and managed internal surveys related to the internal seminars. All WP5 partners actively participated in regular and ad-hoc meetings.

Energy

The Energy pilot representatives completed the Knowledge Space and Graph Data Model. They customised the Energy OpenAIRE Gateway, integrating a previous project, Enermaps. And they drafted the Energy Science Network Graph with several iterations, more will follow. The pilot participated in meetings and training events.

The energy pilot also published a [press release](#) about the ongoing work.

Cancer

The Cancer Research pilot representatives finalised the Knowledge Space and Graph Data Model and customised the OpenAIRE Gateway to meet the community's needs. They collected representative articles for SciLake technical partners to test components and services and provided feedback on prototypes (mainly AvantGraph and BIP! Spaces). They investigated third-party Knowledge Graphs (KGs) for potential integration into the pilot's SKG and developed a first version using a subset of the data. Moreover, they leveraged the SKG in various scenarios, such as Chronic Lymphocytic Leukemia and explored its use for analysing proteins, genes, pathways, diseases, and publications. Finally, the pilot participated in meetings and training events.

Neuroscience

The pilot has customised the Neuroscience OpenAIRE Gateway (<https://ni.openaire.eu>) to ingest related research products, added additional data sources, projects, and subjects and provided input on how data are represented. This gateway is configured to consolidate research outputs from neuroscience repositories and infrastructures. The journals selected as relevant for the neuroscience pilot were based on an established list (<https://research.com/journals-rankings/neuroscience>). The selected repositories were endorsed by the International Neuroinformatics Coordinating Facility - INCF (<https://www.incf.org/infrastructure-portfolio>, filter for 'find data').

By assessing the representation of EBRAINS datasets in the gateway, the pilot has identified aspects that need improvement in the OpenAIRE Graph, such as erroneous duplication or merging of listings and extraction of versioning information. This work has improved the representation of research outputs in the neuroscience gateway and general algorithms in OpenAIRE.

In the first version of the new SKG for neuroscience research, selected metadata from datasets on EBRAINS and other repositories will be transferred onto the AvantGraph instance

associated with the SKG. The data available on EBRAINS are curated and well-annotated with metadata formatted according to the openMINDS framework (RRID:SCR_023173). The data model for the Neuroscience SKG will leverage the openMINDS metadata model to connect datasets with publications. SIRIS' data mining functionalities are implemented to link these to publications from selected journals via openMINDS-controlled terminology.¹ The vision behind this pilot was presented at the [Open Science FAIR](#) in September 2023. Furthermore, a [press release](#) published in June 2024, created with the EBRAINS and SciLake communications teams, was widely distributed on both SciLake and EBRAINS social media channels.

Transportation

Maritime Transportation Research: Pilot representatives finalised the domain knowledge space, provided maritime publications and regulations to technical partners for testing, and participated in joint meetings.

Specifically, the marine transportation team has completed the design of the domain knowledge space and the graph data model. This model includes maritime publications, regulations, research projects, and datasets already integrated into the [OpenAIRE "Transportation Research" Gateway](#). This step was crucial for structuring knowledge and data relevant to the maritime sector. Additionally, sample publications and regulations from the maritime sector were provided to configure and test the SciLake components and services. These publications serve as benchmarks for further refinement and testing of the system, ensuring alignment with the pilot project's objectives. The marine transportation team actively participated in meetings, training sessions, and events to facilitate the smooth development of the Scientific Knowledge Graph and the integration of the pilot project for maritime transportation research. Internal webinars were also organised to provide training on SciLake tools and ensure alignment between provider needs and the pilot's requirements. Finally, the marine transportation team created a [press release](#) to inform the public about the pilot's goals, distributing it on SciLake and CERN social media channels.

CCAM Research: Pilot representatives completed the necessary information for drafting the Knowledge Space and Graph Data Model, customised the "Transportation Research" OpenAIRE Gateway to include CCAM-related projects, and provided related papers for benchmarking. Collected sample publications to assist with configuring and testing SciLake components and services and participated in meetings and training events.

¹ https://github.com/openMetadataInitiative/openMINDS_instances/tree/main/instances/latest/terminologies

6. Pilot evaluation

In this section, we present our plan for evaluating the components and services developed within the project, focusing on scientific experiments and evaluation or testing activities that involve the project pilots and the broader scientific communities they represent. In addition, we describe our approach to assess the project's overall success (mainly piloting) activities.

Background and General Purpose

The SciLake project involves comprehensive evaluation activities designed to assess and validate its outputs. These activities involve analysing all results produced by scientific experimentation and incorporating feedback from the pilot representatives. Naturally, this process engages all partners to ensure thorough technological and use case coverage. As part of this evaluation, the SciLake components and services, particularly those that have undergone significant updates or development during the project's lifetime, are expected to be rigorously tested to ensure their effectiveness and reliability.

According to the project plan, the evaluation and validation of all SciLake technologies began early in the project's lifecycle, starting in its second year (M19 - July 2024). During the initial months, the evaluation process was informal, taking the form of intensive meetings and interactions between the technology providers and pilot representatives. This ad hoc and unstructured approach was intentionally used to inform the design of a more formalised evaluation process, which is expected to be implemented for the remainder of the project.

In the following sections, we describe the overall evaluation methodology and outline the evaluation objectives and the assumptions and limitations underlying the process.

Evaluation Objectives

The primary objectives of the SciLake evaluation process are the following:

- To assess the performance (e.g., efficiency, scalability, accuracy) of key algorithms and models powering central components of the SciLake ecosystem.
- To assess the value of functionalities provided by bundles of SciLake components in real-world case study scenarios inspired by the project pilots while considering end-user satisfaction.

- To evaluate the coverage, accuracy, and functional utility of the domain-specific SKGs created for the project pilots.
- To quantify the overall success of the piloting activities, focus (not exclusively) on their impact on testing the SciLake ecosystem and refining the SKGs.

To achieve these objectives, the SciLake partners have developed a comprehensive, phased, and multi-faceted evaluation process to thoroughly assess the project's components and services. The underlying methodology has been carefully crafted to support achieving the project's goals and is detailed in the following paragraphs.

Methodology

The designed evaluation activities are organised in three phases:

- *Phase 1*) Ad-hoc testing & design (July 2024 - December 2024)
- *Phase 2*) Main evaluation activities (January 2025 - August 2025)
- *Phase 3*) Supplementary evaluation & refinements (September 2025 - December 2025)

During the *first phase*, which has now been completed, the evaluation process was informal and primarily consisted of intensive meetings and interactions between the technology providers and pilot representatives. Specifically, technology providers showcased early prototypes of their technologies and made them available for open testing, enabling pilot representatives to provide valuable feedback. In other instances, they presented mockups or conceptual ideas of functionalities. They engaged in discussions with the pilots to better understand their specific needs and gather insights tailored to the requirements of the respective use cases. This ad hoc and unstructured approach was intentionally adopted to foster open communication and gather valuable stakeholder insights. The feedback and experiences gained during this phase played a crucial role in shaping the design of a more structured and formalised evaluation process, which will be implemented during the second phase to ensure consistency, scalability, and alignment with project objectives. The SciLake partners organised a face-to-face workshop during the third project plenary meeting in Eindhoven in late November 2024 to facilitate productive discussions and finalise the various details of the methodology.

During the *second phase*, the evaluation process will encompass two primary categories of experiments: (a) the developed components, services, and SKGs will be scientifically tested (in many cases) compared with state-of-the-art rivals, and (b) helpful feedback related to the

needs and the experience of the SciLake pilot representatives² will be taken into consideration. The scientific experiments will leverage state-of-the-art ground truth data, methodologies, and metrics appropriate to assess the performance of the developed technologies in terms of efficiency, scalability, accuracy and other properties of interest. On the other hand, the activities to receive targeted feedback will involve pilot representatives getting access to the developed technologies and using them in (structured or unstructured) scenarios related to their unique use cases. The technology providers will determine those scenarios and refine and agree with the partner representatives. Different components and services are expected to be tested in the context of different pilots. It should be noted that these evaluation activities are part of the action plans devised for each pilot (see Annex 8.2). These action plans highlight the main activities expected to take place in the context of each pilot during the creation and enrichment of the respective SKGs and testing or using the developed technologies in practice. In general, the representatives of each pilot are expected to get involved in evaluating the components or services related to their use cases.

The *process's third phase* begins after the main evaluation activities are completed. It focuses on (a) organising campaigns to gather feedback from a broader community of experts, (b) reassessing specific aspects of the performance of some of the SciLake components or services following improvements or modifications implemented, and (c) analysing the results from the piloting and evaluation activities. The campaigns to gather feedback for external experts will be organised by pilot representatives exploiting their domain-specific communication channels and events. Questionnaires, webinars, workshops, and/or special sessions co-organized with established events are expected to occur in this context. Although these activities are primarily planned after August 2025, some may be organized earlier if the SciLake partners deem it feasible and beneficial. The performance aspects for certain SciLake technologies will be reassessed based on the specific needs of the respective technology partners and the availability of pilot representatives (in case their involvement will be necessary). However, it is essential to note that these evaluation tasks will be more limited in scope compared to the extensive activities planned during Phase 2. Finally, the analysis of the results will encompass the entire evaluation process and culminate in preparing the final pilot evaluation report (D5.2), scheduled for completion in December 2025. During this analysis, the partners will also examine various indicators of the success of the piloting activities (e.g., the number of participants, and qualitative and quantitative aspects related to the developed domain-specific SKGs).

² And, in some cases, a broader audience from the respective research communities.

Main Evaluation Activities per Component

In this paragraph, we elaborate in more detail the main activities that we expect to take place during the evaluation process of SciLake. We focus on documenting the evaluation of the primary components and services to be tested. For each³, we outline the specific actions expected to occur as part of the process. The components are presented in lexicographical order.

AvantGraph. AvantGraph is being developed by TUE. AvantGraph will store domain-specific SKGs for each of the pilots. The system will be evaluated based on queries over these SKGs that are of interest to the pilots in the context of graph exploration or analysis scenarios. Those queries will be posed directly to the system using the Lake API or via high-level functionalities of the SciLake value-added services (e.g., BIP! Spaces). For queries, we verify if they are expressible and executable using the AvantGraph Cypher interface, and measure the query execution time. AvantGraph also includes support for running graph algorithms that are beyond the expressive power of traditional graph query languages. Qualitative evaluation of this functionality will be based on user feedback of pilots that have algorithmic workloads to reveal latent knowledge from the graphs (e.g., Cancer pilot representatives have expressed their interest for shortest path and cluster/community detection). For a quantitative assessment, we will evaluate AvantGraph using the industry standard [LDBC Graphalytics benchmark](#). To evaluate performance on a workload tailored to the SciLake project, we will also measure execution time of the PageRank algorithm over the OpenAIRE Graph that is used to compute impact indicators for BIP! Ranker.

Automatic Translation Component. ATHENA RC is developing the automatic translation component. Given its functionality, its primary evaluation will rely on scientific testing without requiring direct feedback from pilot partners. Of course, pilot partners will interact with the component's results, as the translated texts will be indexed and made searchable. This integration is expected to enhance knowledge discovery scenarios by increasing the number of relevant results per query since translated titles, abstracts, or full texts of publications will be naturally linked to specific English keywords, broadening the scope of query results. It is worth mentioning that this component has already been tested using a series of scientific experiments at an earlier stage of the project, with adequate results. More specifically, 1000/1000 sentence pairs were created for development/testing for each pilot. Translations were evaluated over the reference translations using the BLEU, chrF2++, and COMET automated metrics. More details about this experimental series can be found at

³It should be noted that due to their maturity (pre-existing evaluation) or the special nature of the respective task, some of the components and services will not be tested using feedback from the pilots or conducting new scientific experiments.

<https://aclanthology.org/2024.eamt-1.23/>. Until the conclusion of the project, further scientific testing is likely to be conducted to evaluate enhanced versions of the developed models.

BIP! Spaces (Knowledge Discovery Portal). BIP! Spaces is the Web portal service that implements the main knowledge discovery functionalities and is being developed by ATHENA RC. Developed as part of the BIP! ecosystem, it facilitates the exploration of a vast repository of scholarly outputs, such as publications, datasets, software, and other research products, offering advanced citation-based impact indicators provided by the citation-based impact analyser for effectively ranking or filtering search results. BIP! Spaces also enriches the search results with domain-specific annotations derived from specialised knowledge graphs, providing experts with enhanced context. Given their entities of interest, these annotations aim to make information more relevant and actionable for expert users across the various domains. To evaluate them, a feedback mechanism will be implemented where users rate annotations using a simple and intuitive mechanism (e.g., like or dislike). In general, the user satisfaction assessment activities for this service will also be used to provide relevant feedback for testing and evaluating various other components of which the outputs are being displayed. A survey can then be conducted among the pilot users, aggregating feedback data (e.g., the frequency of likes and dislikes for specific annotations), analysing their patterns, and refining the annotations to achieve greater usefulness. Due to the availability of graph contents via the UI elements of BIP! Spaces, the evaluation activities for this service will be also useful for the assessment of the respective SciLake components that offer various enrichments for the graphs.

Citation-based impact analyser. ATHENA RC and SIRIS are developing the citation-based impact analyser, a tool designed to compute a set of citation-based indicators over a citation network. A large part of this component is based on ATHENA RC's BIP! Ranker (<https://github.com/athenarc/Bip-Ranker>) software library that focuses on evaluating various dimensions of scientific impact, including popularity (current impact), influence (overall impact), and impulse (initial momentum). In the context of the project this library is extended to consider additional information for the citations (e.g., additional citation context) for the calculation of impact indicators. In addition, special attention is given to developing tailored indicators for datasets, ensuring their unique characteristics and contribution are accurately measured. This activity is related to the creation of a sub-component to be co-developed by ATHENA RC and SIRIS. This leverages the outputs from other components of the project, such as the SciNoBo RAA tool, in order to take informal citations to datasets/software, as well as their intent, into consideration. Several aspects related to the effectiveness of the analyser have been evaluated in older, published scientific experiments. New experiments of similar nature are expected to be conducted in the context of the project giving special focus on the new functionalities, like those related to the dataset impact analysis. Due to the fact that there

exists no ground truth for the impact of these research objects, the evaluation will be carried out in a qualitative manner, gathering the user experience of the pilot communities when using the newly-proposed dataset indicators as a knowledge discovery aid. The outputs of the various sub-components of the impact analyser are being integrated into the SciLake SKGs to facilitate knowledge discovery and other similar tasks. Because of that, their value will be assessed in the context of user satisfaction surveys that will involve the Lake API, BIP! Spaces, and SciNoBo assistant.

Domain-specific entity recognition component. This component is being developed by SIRIS and OPIX. There are two similar but distinct software packages, each focusing on different pilot cases. Its aim is to enrich the SKGs by identifying and disambiguating mentions of entities that are relevant to the specific research communities in a tailored fashion, using the texts from publications and/or regulatory/legal documents. The evaluation will be carried out during development using standard metrics (F1 scores based on strict matches) using a combination of existing gold-standard datasets and pre-annotated data validated by the pilot representatives, depending on the specific entities being addressed. In addition to this, the usefulness of the tool will be assessed qualitatively by gathering user feedback from the pilot communities using a small set of documents and the document-entity mapping extracted automatically.

DoStRe (Article Segmentation Component). DoStRe is the article segmentation component and is being developed by DFKI. The tool has been tested in the past using standard measures (like accuracy, precision, recall) to assess the effectiveness in section classification and similar experiments may be conducted in the context of the project to evaluate the contributions of developed improvements and exploit pilot-generated data and feedback. In general, the tool is expected to be bundled together with other tools (e.g., SciNoBo CA tool, entity recognition components) in the context of particular pilot-related use cases, hence indirect evaluation of its effectiveness will also happen based on this.

KG creation assistant. This tool refers to an array of SciLake components that can be used to assist domain experts with the creation of SKGs. The four main components are *R2PG-DM*, *ProGGD*, *GDDMiner* (all three by TUE), and the *Affiliation disambiguation component* (by SIRIS).

R2PG-DM is a component that aids in creating property graphs from relational databases by supporting a direct mapping that follows a natural and logical translation of the relevant properties. *ProGGD* is a system that discovers Graph Generating Dependencies (GGDs) – a novel class of expressive integrity constraints in property graphs that capture topological patterns, property value similarities, and differences in graph data – subsuming other graph dependencies to address various data quality and management challenges. Lastly, *GDDMiner*

tackles the problem of discovering Graph Differential Dependencies (GDDs) by identifying a non-redundant and concise set of GDDs that hold in a given property graph, efficiently producing a minimal cover of valid dependencies.

Due to the nature of the task that these components are responsible for (i.e., transformation of relational databases into labeled property graphs) and the fact that the used version of the component has been tested in the past, there is no need for additional evaluation in the context of the SciLake project.

Regarding the *Affiliation disambiguation component*, the evaluation is planned to be based on both intrinsic scientific testing to assess its performance using standard metrics like F1 scores based on strict matches. User feedback targets the utility and relevance of the automatically identified disambiguated institutions extracted from a small, predefined set of publications implementing a qualitative approach, which may be more lenient than the computed metrics; eg., imperfect matches requiring a minimal amount of curation may still be considered useful.

PDFfetcher. PDFfetcher is a tool developed by OpenAIRE. Due to its nature, its value can be assessed based on the number of full-text manuscripts it will retrieve by the end of the project (see Section “Evaluation of project activities success” and KPI K4). The pilots will interact with the tool’s outputs, as it is integrated with text mining and natural language processing (NLP) components that introduce additional nodes into the domain-specific SKGs. Consequently, the success of activities involving these components depends on the performance of the PDFfetcher. Its effectiveness is expected to influence the level of enrichment achieved within the SKGs directly.

Reproducibility & Replicability Assistant. This service refers to a series of UI elements that collect information from the SciNoBo RRA and the SciLake SKGs and present it to the users in an intuitive way to inform them about the reproducibility of works of interest. BIP! Spaces implements some of those elements and it will be investigated to implement similar elements in the SciNoBo assistant (used for the evaluation of SciNoBo suite components, see below). The evaluation of the service will be on the basis of user feedback and satisfaction that will happen organically and in parallel with the evaluation of other functionalities of the involved components during the piloting activities.

Research Object Link Recommendation component. This tool refers to an array of SciLake components that can be used to recommend missing links in KGs. The two main components are *SciNeM* (by ATHENA RC) and *SHINER* (by TUE).

SciNeM is a powerful Knowledge Graph (KG) analysis tool that focuses on metapath-based analysis, by leveraging the graph structure to uncover missing links representing latent and complex knowledge encoded in the graph structure.

SHiner, on the other hand, is a tool that can be used to recommend missing links among graph entities based on a predefined set of Graph Generating Dependencies (GDDs). An indicative example is to identify missing edges (i.e., links) between research products and projects in the OpenAIRE Graph.

SciNeM and *SHiner* are both tools designed to identify links between nodes in a Knowledge Graph, making them comparable in their functionality. Both have been previously evaluated in relevant publications for their scalability and effectiveness. However, within the scope of this project, the focus will shift to the user perspective, evaluating user satisfaction for the outputs of those tools through BIP! Spaces, SciNoBo assistant, or even the Lake API.

SciNoBo suite. SciNoBo suite refers to an array of tools provided by ATHENA RC: the *CA tool*, the *FoS component (taxonomy & classifier)*, the *RAA tool*, and the *SDG classifier*. All these tools are expected to be tested on the basis of intrinsic, scientific experiments, as well as user satisfaction. Apart from the scientific experiments, the SciNoBo team will also leverage the SciNoBo Assistant to further evaluate the tools and assess user-satisfaction. The SciNoBo Assistant, is a platform that provides an interactive way for evaluating the effectiveness of the tools under evaluation, along with user-satisfaction. Researchers will be able to interact with the Assistant through natural language queries to explore the tools' outputs giving ratings and qualitative comments, offering insights into precision, granularity, and utility. This feedback loop will help the identification of areas for improvements in each of the tools. In the following paragraphs we provide details for each of the tools under evaluation.

The *CA tool* is capable of performing citation analysis to classify citations according to their intent (generic/reuse/comparison), polarity (supporting/neutral/refuting), and semantics (claim/methodology/results/artefact). An intrinsic, scientific testing will evaluate the tool using typical measures for the task (e.g., PRF scores). An initial human-curated dataset of approximately 1200 instances has been created to scientifically evaluate the tool. PRF scores will be calculated across the intent, polarity and semantics. Furthermore, PRF scores will be calculated for the subcategories of the intent (e.g. generic), polarity (e.g. supporting) and semantics (e.g. claim). Additionally, a number of judgments on classified citations will be collected from pilot representatives, as well as user-satisfaction and qualitative comments, based on their interaction with the integrated feedback mechanism of the SciNoBo Assistant. The respective experiment will involve pilot representatives and external experts from the same domains (where possible).

The *FoS component* comprises an FoS taxonomy and a classifier. For the taxonomy, scientific experiments are expected to be used to assess its coverage (e.g., the percentage of relevant fields compared to a reference standard like EurosciVoc and Scopus), its field density (e.g. measuring the average number of subfields per parent field across levels and compare them with a reference standard) and its comprehensiveness utilizing approaches using Large Language Models (LLMs) as LLM-as-a-judge⁴. An initial evaluation and comparison among different taxonomies including the SciNoBo taxonomy is outlined [here](#)⁵. For the classifier, a comparison against a Deep Learning model has already been conducted at L1-L3, calculating MACRO-F1, WEIGHTED-MACRO-F1, MICRO-F1 ([Link](#)). Similar experiments (measuring Precision, Recall, F1 Score for the effectiveness) may be expected to re-evaluate changes and improvements to the classifier at the same or different taxonomy levels. The interaction of the pilot representatives with the tools' results through the feedback mechanism of the SciNoBo Assistant will assess user satisfaction. A similar approach is expected for the evaluation of the *SDG classifier*, which categorises publications based on the Sustainable Development Goals (SDGs). It is a multi-label classifier that assigns multiple labels to a publication, allowing researchers to align their work with specific SDGs.

The *RAA tool*, is capable of performing research artefact analysis (RAA) to extract mentions of both named and unnamed research artefacts (RAs) (e.g., datasets, software) from scientific text, along with specific metadata attributes for these artefacts, such as Name, Version, License, and URL. It also classifies these artefacts based on their Usage and Provenance, identifying whether the artefacts were created or used by the authors of the scientific text. An intrinsic, scientific evaluation of the tool will use standard metrics such as the F1 score, which measures the successful identification of valid RA mentions or the presence of metadata attributes (e.g., Name, License, Version, and URL). For Usage and Provenance, the F1 score reflects the accurate identification of the authors' interaction with the RA. Additionally, the exact match (EM) score assesses the exact lowercase match of metadata text within a provided snippet, while the lenient match (LM) score evaluates whether the model's lowercase prediction is contained within the ground truth or vice versa. A scientific evaluation of the tool has already been conducted and is outlined in the publication [Empowering Knowledge Discovery from Scientific Literature: A novel approach to Research Artifact Analysis](#). Finally, following the same approach with the abovementioned tools, constructive user feedback will be collected through the interaction of the pilot representatives with the tools' results through the feedback mechanism of the SciNoBo Assistant.

⁴ Code and results are and will be made available in the following repository: [validator-experiment](#)

⁵ The paper cites and preliminarily evaluates our SciNoBo taxonomy (refer to the name OpenAIRE's Field of Science taxonomy in the paper for the results).

Topic analysis component. Topic analysis component is being developed by ATHENA RC. This tool provides valuable insights for each dominant (i.e., frequently appearing) topic in the search results of BIP! Spaces. Specifically, users can explore their trends over the years in the form of intuitive visualisations, enabling them to understand emerging trends and assess the impact of specific topics. The tool has been organically integrated into the BIP! Spaces component, hence its evaluation will be based on the user satisfaction assessment that is described for this component. More specifically, to evaluate the effectiveness of the provided insights, the same feedback mechanism proposed for BIP! Spaces (i.e., a like/dislike system for user input) will be used.

Additional Testing by External Experts

In addition to the evaluation activities mentioned earlier, the project partners plan to conduct outreach campaigns to engage external experts from the research communities involved in the pilots to gather additional feedback (mainly during phase 3 of evaluation but often at earlier stages, as well). Since the participation of external experts will be voluntary, these activities are designed to be lightweight. Feedback is expected to take the form of informal input during presentations and/or hands-on workshops centred on SciLake technologies and brief, targeted questionnaires to collect specific insights.

The specific focus of these testing activities will be refined by the SciLake partners closer to organising the respective events, allowing flexibility to address targeted topics identified during the main evaluation activities. The pilot partners will take responsibility for identifying the audience by utilizing their domain-specific communication networks and leveraging opportunities to organize project-related events alongside well-established conferences. In the following paragraphs, we outline the planning conducted by each pilot.

Energy

The HES-SO team participating in the SciLake project plans to deliver an internal presentation at HES-SO in the summer of 2025 after completing the leading testing and evaluation activities. This presentation will introduce the piloted SciLake technologies to the organisation's staff and gather feedback from a broader group of internal experts. Additionally, during the same period, the team plans to present the same technologies in the context of the [project OpenGiS4ET](#), in which HES-SO also curates [open energy data](#). Finally, in the fall of 2025, the team intends to showcase the SciLake technologies to focus groups and present their work through a keynote presentation at the annual [ECMP conference](#). The exact date (end of October or November) needs yet to be defined for 2025. On all these occasions,

the HES-SO team will aim to gather targeted feedback on the technologies and relay it to the respective technology-providing partners, as needed.

Based on the previous, it is clear that the primary focus of these actions will be to engage stakeholders from the broader community of researchers in the energy research field, as well as policymakers and industry professionals interested in this domain.

Cancer

The pilot representatives have already established connections with experts in cancer research from Uppsala University and the Department of Oncology-Pathology at Karolinska Institutet, with plans to involve additional research groups. A targeted dissemination event (such as a webinar or an on-site workshop) is scheduled for June 2025 to demonstrate and share the Cancer SKG. During the workshop, researchers will engage in hands-on testing sessions and provide structured feedback through practical exercises. The respective sessions will involve completing real research tasks using the developed Cancer SKG, either directly or via the specialised user interface designed by the SciLake partners, to evaluate the completeness of the SKG and the value of the respective technologies in different realistic scenarios. Additionally, pilot representatives will also consider distributing a specifically designed questionnaire to evaluate user satisfaction. The collected insights will be compiled into a feedback report to guide further improvements.

Additionally, the pilot representatives will explore opportunities to leverage well-established conferences and events in the Cancer Research domain (e.g., ECCB2025 <https://www.iscb.org/ismbecb2025/home>) to organize specialized sessions or presentations. This approach aims to engage a broader community of external experts and gather valuable additional feedback.

Based on the previous, it is evident that the primary focus of the designed actions will be to engage researchers from the Cancer research domain.

Neuroscience

The Neuroscience pilot will showcase the pilot to the community of EBRAINS users and collect feedback about the functionality and utility of the metrics and KG enrichments provided. EBRAINS brings together a diverse group of scientists and researchers from various fields within neuroscience, such as computational neuroscience, cognitive neuroscience, neuroinformatics, and clinical neuroscience. The community can be reached through the EBRAINS Community platform (<https://community.ebrains.eu/>) and via various mailing lists

and newsletters. Among other activities, researchers will be invited to participate in hands-on testing sessions and provide structured feedback through questionnaires. In addition to this, the SciLake pilot will be presented at the FENS Regional Meeting 2025 (FRM 2025), taking place June 16-19, 2025, in Oslo, Norway (<https://frm2025oslo.no/>).

Transportation

Maritime Transportation Research. The Maritime Transportation Research pilot team will be responsible for evaluating the accuracy and relevance of SciLake tools in processing maritime publications and regulations to generate domain knowledge, as well as for interlinking data from various maritime sources to create a coherent domain-specific SKG. Additionally, the usability of the knowledge graph and its associated technologies will be tested with maritime stakeholders, focusing on how effectively they can interact with SciLake services to navigate and derive insights from the maritime knowledge graph.

The usability testing process will begin with the recruitment of participants, followed by the development of test plans and the setup of the testing environment. The testing and evaluation activities will start immediately afterwards, proceeding through iterative phases of testing and result analysis. Initial iterations will prioritize feedback from pilot representatives (as part of main evaluation activities), with the gradual involvement of recruited external experts as the process progresses.

The process will aim to involve several categories of stakeholders within the maritime transportation research community, including technology providers, research bodies (i.e., association, institution, society, institute, fraternity, chamber, group, board), and funding agencies. Adjustments to tools such as entity recognition, automatic translation and GGD-based interlinking will be made based on user feedback. To ensure effective engagement and dissemination of SCILAKE services, several actions are proposed: the organization of a workshop at the 12th International Congress of Transport Research (ICTR 2025) to present SCILAKE tools to a wide audience; the presentation of SCILAKE services at the WATERBORNE Technology Platform, leveraging CERTH membership to connect with innovators in the maritime transport sector; and the organization of a workshop with the Piraeus Chamber of Commerce and Industry to present SCILAKE tools to representatives of the local maritime industry in Piraeus. These actions aim to foster collaboration, gather insights and evaluate the SCILAKE tools to better meet the needs of the maritime research and industry communities.

CCAM Transportation Research. The main target audience for the CCAM Transportation Research evaluation consists of researchers from the Research Institute of Computer and Communication Systems (ICCS). Apart from members of the CCAM sub-team of i-Sense, which

is the ICCS Department involved in the project, researchers within other sub-groups of ICCS and other Academic Institutes will be approached to use and evaluate the SKG and all relative services. Those external experts will be involved at an early stage taking part in the iterative phased process of testing and analysing.

The dissemination Department of i-Sense sub-group has listed several CCAM-related transport events where individuals will be approached (indicatively ITS Europe, TRA 2025, CCAM related projects internal workshops) and acknowledged on SciLake's scope. Moreover, the CCAM Transportation pilot will co-organize the 12th International Congress of Transport Research (ICTR 2025). Feedback from the target audience from all above events will be guided through questionnaires on the functionality of each service, the general feeling of the queries, the effectiveness of the desired queries to the SKG, the relateness of the returned results for the CCAM transportation sub-community, etc.

Evaluation of project activities' success

In addition to the previous, SciLake partners have determined a set of Key Performance Indicators (KPIs) to assess the overall success of the project activities (mainly piloting but also developing). A set of relevant KPIs have been identified in the project proposal and have been updated in the context of Deliverable D1.1⁶ ("Initial service requirements"):

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

⁶ SciLake D1.1: <https://zenodo.org/records/14335008>

- Initial: 0
- Target: $\geq 10k$

These partners will use SciLake KPIs to monitor the success of SciLake services in addressing the key pilot needs in the context of their use cases.

In addition to the previous KPIs, all pilots will measure the number of external experts involved in the SciLake evaluation activities as an indicator for the success of the respective engagement activities.

DRAFT

7. Conclusions

The WP5 initiative has played a pivotal role in advancing domain-specific Scientific Knowledge Graphs (SKGs) and related services within the SciLake project. These efforts aim to improve research accessibility and drive innovation across diverse scientific fields. Employing an agile, iterative roadmap, WP5 has effectively managed key phases, including requirements elicitation, preparation, pilot execution, and evaluation, ensuring responsiveness to evolving research demands and stakeholder feedback.

The project encompasses four primary pilots in Neuroscience, Transportation, Cancer, and Energy research, each addressing critical challenges through tailored SKGs. The Energy pilot focuses on regional energy planning and sustainable, carbon-neutral solutions. The Cancer pilot seeks to enhance personalized medicine for Chronic Lymphocytic Leukemia by identifying biomarkers and refining treatments. Neuroscience emphasizes data curation and the integration of datasets with publications, while Transportation targets Maritime and Cooperative Connected Automated Mobility (CCAM) to develop unified SKGs that enhance decision-making and innovation.

Collaboration between pilot teams and technology developers has been central to WP5's success. Through internal seminars, workshops, and iterative reviews, partners aligned their efforts, shared expertise, and refined SciLake tools. These activities facilitated the identification of domain-specific knowledge spaces, critical entities, and relationships, enabling the creation of high-quality prototypes. Regular feedback and training sessions ensured that technological solutions were closely aligned with the distinct needs of each domain.

A robust evaluation framework underpins the SciLake project, incorporating informal feedback, structured testing, and outreach to external experts. Key Performance Indicators (KPIs), such as the number of SKGs created and stakeholder engagement levels, measure success and guide refinements. This approach ensures the technologies are practical, effective, and valuable in real-world applications, while continually enhancing the SciLake ecosystem.

The project remains on track to deliver a beta version of its SKGs and services by June 2024, marking a significant milestone in scientific knowledge management. Guided by Open Science principles, SciLake's iterative approach ensures continuous improvement and alignment with its mission to support research accessibility and innovation.

8. Annexes

8.1. Press releases

All press releases were published on LinkedIn by the respective partner institution, and they are published [here](#) on the Scilake website. Hereunder it follows the full text press release per each pilot.

- [Transportation Maritime](#)
- [Transportation CCAM](#)
- [Cancer](#)
- [Energy](#)
- [Neuroscience](#)

Press Release - SciLake Maritime Transport Pilot



PRESS RELEASE

Comprehensive understanding of the complex and diverse aspects of research in maritime transport

SciLake's "Transport pilot – Pioneering Knowledge Graph in Maritime Transport Research"

Piraeus, Greece, May 13, 2024

The SciLake project [1] announces the first scenario in the transportation pilot case focused on maritime transport research led by researchers of the Hellenic Institute of Transport (HIT) at the Centre for Research and Technology Hellas (CERTH), Greece. SciLake aims to leverage the [OpenAIRE](#) ecosystem and develop innovative knowledge management and discovery solutions to provide domain-specific services for the transportation research community. These services will be based on a Scientific Knowledge Graph focused on transportation research and will offer smart knowledge discovery capabilities leveraging impact and reproducibility indicators serving several categories of stakeholders within the transport research community, including technology providers, research bodies (i.e., association, institution, society, institute, fraternity, chamber, group, board), and funding agencies.

Regarding Technology Providers, this scenario aims to offer validator services that automatically verify the compliance of industry's predefined guidelines. Additionally, full-text mining algorithms utilise the full-texts of publications to enhance metadata records by linking them to various research-related entities such as research projects, publications, datasets, affiliations, journals/conferences, funding agencies and research fields/topics and to domain-specific ones, such as regulations and vessel types.

Press Release - SciLake Maritime Transport Pilot

For Research Bodies, the services aim to facilitate knowledge discovery and support the adoption of Open Science publishing practices and monitor the implementation of Open Science practices by their researchers. Using SciLake services, users can access content reports and project lists. The content reports provide information about the research outcomes associated with the bodies, while the project lists outline the projects the research body is involved in and their respective research outcomes.

Lastly, Funding Agencies, such as national and international research funding organisations, can access services to monitor the impact of their funding within the maritime transport research community.

Moreover, the maritime transport research scenario will be built around the main waterborne visions [2] (such as green and clean waterborne transport, safe and secure waterborne transport; port operation; etc.) including re-usability re-posts for publications and relevant datasets.

References

[1] <https://scilake.eu/>

[2] <https://www.waterborne.eu/>

About CERTH/HIT

CERTH - Centre for Research and Technology Hellas (www.certh.gr) is the largest research centre in Greece with the mission to carry out basic and applied research with special emphasis on exploiting research results and developing new products and services with industrial, economic and social impact. CERTH is currently ranked in the 10th position among the EU Research Centres in terms of H2020 Net EU Contribution and 1st in Greece. It comprises 5 main research institutes including the Hellenic Institute of Transport (HIT). HIT (www.imet.gr) is the Greek National Institute for promoting Transport Research and Policy Support. It focuses on applied research in all fields and modes of Transport, providing input for policy formulation, documentation of major trends and impacts, formulation of operational rules and procedures, and improvement of the operation and management in Transport.

About SciLake

SciLake is a project funded by the European Union's Horizon Europe program. The project aims to create a seamless integration between domain knowledge and open Scientific Knowledge Graphs, while also developing useful added-value services for

Press Release - SciLake Maritime Transport Pilot

specific research areas. The ultimate goal is to empower researchers and foster a more interconnected and efficient scientific community. SciLake brings together a competent consortium of 13 partners from 9 different countries. The consortium consists of partners with expertise in knowledge management and discovery, as well as experts from Neuroscience, Cancer, Transportation, and Energy research, who are involved in piloting activities.

Contacts:

Afroditi Anagnostopoulou - a.anagnostopoulou@certh.gr

Xenophon Kitsios - xkitsios@certh.gr

Press Release - SciLake Transport CCAM Pilot



PRESS RELEASE

Using SKGs in the emerging sector of Cooperative Connected Automated Mobility (CCAM)

SciLake's "Transport pilot – 2nd Scenario: Guiding CCAM Research through Scientific Knowledge Graphs".

Athens, Greece, May 14, 2024

The Transportation pilot of the SciLake project [1] announces its second scenario, which pertains to research in the emerging sector of Cooperative Connected Automated Mobility (CCAM), led by ICCS researchers in Greece. Knowledge from existing research related to co-funded European projects around CCAM will be used to enrich the previously developed Transportation Scientific Knowledge Graph (SKG) [2] of the OpenAIRE community [3] and implement the project's advancements.

The current version of the Transportation SKG has been rendered obsolete due to its outdated information regarding CCAM. With the rapid evolution of the sector, it is crucial for researchers to keep up with new trends, definitions, key terms and any other forms of expert knowledge. To address this gap, a detailed taxonomy around CCAM has been identified from another EU funded project called SINFONICA, which is publicly available under a relevant deliverable [4].

The next step is to analyze this taxonomy and cooperate with technical partners from the SciLake project to identify valuable enrichments for the SKG based on entities mentioned in the taxonomy and on patterns of importance related to CCAM in scientific metadata and texts. Experts from the domain will then be in place to examine the results, provide feedback, and identify new domain knowledge. As a result, new

Press Release - SciLake Transport CCAM Pilot

datasets are expected to be unveiled, significant papers to be identified and current research trends to be revealed.

In parallel, an existing Knowledge Base [5] provides valuable data to be incorporated into the project's developed models. Covering a wide range of aspects around CCAM, from key enabling technologies to legal documentation and general thematic categories, this resource will be appropriately modified to be ready for use by the project's models. The ultimate goal is to fine tune the targeted extracted results to assist researchers in their work on CCAM.

References

- [1] <https://scilake.eu/>
- [2] <https://beopen.openaire.eu/>
- [3] <https://www.openaire.eu/>
- [4] https://sinfonica.eu/wp-content/uploads/2023/07/D1.3-Understanding-the-Gap-of-CCAM-solutions-deployment_v1.0.pdf
- [5] <https://www.connectedautomateddriving.eu/>

About ICCS/I-Sense

I-SENSE (<https://i-sense.iccs.gr/>) is one of the Research Groups of the Institute of Communication and Computer Systems (ICCS) of the National Technical University of Athens. I-SENSE Group is very active in a number of Scientific and Research Areas with main Application Areas the Intelligent Transportation Systems, Virtual Environments, Assistive Technologies, Smart Integrated Systems – Sensors, Communication, Platforms.

About SciLake

SciLake is a project funded by the European Union's Horizon Europe program. The project aims to create a seamless integration between domain knowledge and open Scientific Knowledge Graphs, while also developing useful added-value services for specific research areas. The ultimate goal is to empower researchers and foster a more interconnected and efficient scientific community. SciLake brings together a competent consortium of 13 partners from 9 different countries. The consortium consists of partners with expertise in knowledge management and discovery, as well as experts

Press Release - SciLake Transport CCAM Pilot

from the fields of Neuroscience, Cancer, Transportation, and Energy research, who are involved in piloting activities.

Contact:

Athanasios Ballis - athanasios.ballis@iccs.gr

###

Press Release - SciLake Cancer Pilot



PRESS RELEASE

Connecting the Dots between Cancer Data Networks: The SciLake Cancer Knowledge Graph

April 30th, 2024

Researchers from the Centre for Research and Technology in Greece and the Department of Molecular Medicine and Surgery, Karolinska Institutet in Sweden set up a collaboration to make publicly available resources in biology and cancer more accessible to the research community. The teams are developing a first-of-its-kind cancer knowledge graph (KG), as part of the SciLake Project involving partners across Europe.

The focus of this pilot project is to harness the power of SciLake services to aid in identifying biomarkers essential for personalised treatment and care, a crucial step towards the realisation of precision medicine. The case study of interest is Chronic Lymphocytic Leukemia (CLL), the most common adult leukaemia, characterised by a highly heterogeneous clinical course. The goal of this pilot is to gain a better understanding of this heterogeneity by leveraging information from the CLL knowledge graph (KG), a cancer-specific KG that can effectively model the respective knowledge space using defined data sources. CLL KG will combine existing biomedical KGs like the Clinical Knowledge Graph with text mining and entity recognition services to reveal connections between clinical covariates, genes, proteins, metabolites, mutations, drugs, scientific literature, biomedical datasets, software and research consortia.

The potential applications of this cancer 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 use the KG to uncover latent knowledge using more advanced graph mining algorithms and machine

Press Release - SciLake Cancer Pilot

learning approaches. In the context of precision medicine, the CCL-KG will 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 teams have been diligently searching for existing knowledge graphs in the context of cancer. Recognising the ongoing need for an updated, cancer-specific knowledge graph, they have invested significant time in sourcing more general knowledge graphs from biology and precision medicine that will serve as a starting point. The challenge ahead lies in the effective integration and enrichment of these graphs to create a comprehensive cancer graph.

However, the potential benefits far outweigh the challenges. A successful cancer KG will be a significant asset for the wider cancer community. This tool, leveraging graph-based algorithms, user-friendly analytics, diverse clinical-biological databases, and drug-specific information, will empower scientists in their research.

About Karolinska Institutet

Karolinska Institutet is one of the world's leading medical universities. Its vision is to advance knowledge about life and strive towards better health for all. Karolinska Institutet accounts for the single largest share of all academic medical research conducted in Sweden and offers the country's broadest range of education in medicine and health sciences. The Nobel Assembly at Karolinska Institutet selects the Nobel laureates in Physiology or Medicine.

About CERTH

The Centre for Research & Technology HELLAS (CERTH) is a leading research centre in Greece conducting specialised basic and applied research and offering high quality services in Life Sciences. The Institute of Applied Biosciences at CERTH (INAB|CERTH) is dedicated to the promotion and execution of applied biosciences. Research at INAB|CERTH revolves around molecular medicine, biomedical informatics, and the development and evaluation of eHealth, mainly in chronic diseases.

Press Release - SciLake Cancer Pilot

About SciLake

SciLake is a project funded by the European Union's Horizon Europe program. The project aims to create a seamless integration between domain knowledge and open Scientific Knowledge Graphs while also developing useful added-value services for specific research areas. The ultimate goal is to empower researchers and foster a more interconnected and efficient scientific community. SciLake brings together a competent consortium of 13 partners from 9 different countries. The consortium consists of partners with expertise in knowledge management and discovery, as well as experts from Neuroscience, Cancer, Transportation, and Energy research, who are involved in piloting activities.

Contacts:

Dania Machlab	daniamachlab@ki.se
Leily Rabbani	leily.rabbani@ki.se
George Gavriilidis	ggavri@certh.gr
Konstantinos Kardamiliotis	k.kardamiliotis@certh.gr
Fotis Psomopoulos	fpsom@certh.gr
Vasileios Vasileiou	vasileioubill95@certh.gr

###

Press Release - SciLake Energy Pilot



PRESS RELEASE

Sustainable regions require regional energy planning

SciLake's "Energy pilot" launches "Regional Energy Planning Pilot"

Sion, Switzerland, April 26, 2024

In a strategic move towards sustainable energy solutions, the SciLake project [1] announces the refinement of its Energy Pilot into the Regional Energy Planning (REP) Pilot. This initiative builds a research ecosystem that seamlessly contextualises, interconnects, and makes scientific knowledge interoperable and accessible based on the know-how of the HES-SO Valais-Wallis. The Regional Energy Transition includes specific characteristics tailored to diverse geographical contexts.

Regional systems are often linked to specific forms of governance: a higher level of governance defines the overall goal that is then further contextualised at the regional level into actionable plans before being implemented by local governments such as cities and smaller communities. The context includes the secure, resilient, and carbon-free regional energy supply, leveraging the flexibility of locally and regionally available energy sources. The focus is on literature aligning with local and regional requirements about generation, demand, and sustainability while optimising the use of infrastructures and resources. Energy sobriety is a critical driver in significantly reducing energy consumption.

The REP is managed by the HES-SO Valais/Wallis team of "sustainable energy and territory" led by Prof. Jakob Rager within the Institute of Energy and Environment at the

Press Release - SciLake Energy Pilot

School of Engineering. The team gathers competencies related to energy planning, project management and environmental engineering and actively works on energy planning research and applied research projects at local, regional, national and European scales.

HES-SO's "sustainable energy and territory" team manages the REP platform. Prof. Jakob Rager, leading the team, explains: "Our activities focus on our holistic system analysis to derive sustainable policies." The team builds energy planning, project management, and environmental engineering competencies. It actively supports actors on multiple levels, from local to regional or national, to develop more actions towards a sustainable world. It combines European scales related to energy planning research and applied research.

As part of the SciLake project, the REP, one of the four pilots, uses cutting-edge tools to build scientific knowledge graphs (SKGs). These SKGs connect data and capture expert knowledge, facilitating the collection, organisation, and retrieval of heterogeneous scholarly content.

The pilot aims to leverage SciLake tools for semantic modelling, maintenance of ontologies, extraction of knowledge from unstructured raw data, and graph mining to create a knowledge graph that will facilitate the discovery of knowledge sharing in the respective field. To this end, advanced knowledge discovery approaches that exploit impact and reproducibility indicators for research products (e.g., publications and datasets) will also be used. The EnerMaps Open Data Management Tool [2] will notably be enhanced to support the Regional Energy Transition initiative.

By aligning with regional energy planning, SciLake's Energy Pilot REP marks a significant step towards sustainable regions. By integrating advanced tools and methodologies, this pilot addresses energy challenges and contributes to a broader research ecosystem that enhances the accessibility and interoperability of scientific knowledge. The journey towards regional energy sustainability begins here in Sion, Switzerland.

References: [1] <https://doi.org/10.3030/101058573> ; [2] <https://enermaps.openaire.eu/> .



scilake.eu



[SciLake project](#)

Supporting



Funded by
the European Union

Press Release - SciLake Energy Pilot

About HES-SO

The HES-SO is the largest university of applied sciences (UAS) in Switzerland and the country's third largest higher education institution, with more than 21'000 students and a broad network of schools in 7 cantons. HES-SO is organised into six faculties: Design & Visual Arts, Business Management and Services, Engineering & Architecture, Music & Performing Arts, Health and Social Work, and education and research are oriented towards practical applications in the same areas. HES-SO offers many education programs: 46 Bachelor's degrees, 26 Master's degrees and 301 continuing education courses. It employs more than 17'000 collaborators, with approx. 960 FTE dedicated to R&D activities. Firmly anchored in the regional economy, HES-SO collaborates closely with SMEs, and its R&D also extends to certain aspects of industrial-scale production. HES-SO undertakes research projects with various partners in Switzerland and abroad. HES-SO has been involved in European Framework Programs since 1998. In Horizon 2020, we participated in 45 collaborative projects, including six as coordinator and covering the program's three Pillars. We are participating in 15 projects in Horizon Europe as an associated partner.

About SciLake

SciLake is a project funded by the European Union's Horizon Europe program (grant No. 101058573). The project aims to seamlessly integrate domain knowledge and open Scientific Knowledge Graphs while developing valuable added-value services for specific research areas. The ultimate goal is to empower researchers and foster a more interconnected and efficient scientific community. SciLake brings together a competent consortium of 13 partners from 9 different countries. The consortium comprises partners with expertise in knowledge management and discovery and experts from Neuroscience, Cancer, Transportation, and Energy research involved in piloting activities.

Contacts:

Alejandro Pena-Bello: alejandro.penabello@hevs.ch

Andrea Salmi: andrea.salmi@hevs.ch

Jakob Rager: jakob.rager@hevs.ch

###

Press Release - SciLake Neuroscience Pilot



PRESS RELEASE

EBRAINS and SciLake collaborate to improve data services for neuroscience research

June 4th, 2024

The EBRAINS research infrastructure recently announced a strategic collaboration on a neuroscience pilot project with SciLake, a EU-funded project that supports open science. This partnership aims to enhance neuroscience research by interlinking the EBRAINS data network (Knowledge Graph) with SciLake's services. The joint effort focuses on improving data linkage with scientific articles and identifying trends and gaps in neuroscience research through advanced analytics of data availability and usage.

SciLake assists scientific communities in constructing advanced data networks mapping data from multiple sources and creating connections between different entities (researchers, organisations, funders, and facilities) of a given topic.

Neuroscience is a complex field studying the basic properties of the brain, disease-related changes, and application of brain-inspired sciences. A significant challenge in this field is how to handle the large diversity of data sources and formats, which limits the integration of data and reproducibility of research outputs. EBRAINS offers a research infrastructure that is tailored for integrating heterogeneous data and metadata. The collaboration between EBRAINS and SciLake will connect the EBRAINS data network to the wider pool of publication data in the OpenAire Graph, enabling analysis of information from various resources and more efficient tracking of progress and trends. The advanced text mining functionalities of OpenAire will greatly expand the possibilities for data providers and users to analyze the uptake and relevance of data organized in the EBRAINS Knowledge Graph.

Press Release - SciLake Neuroscience Pilot

The collaboration anticipates improved data workflows, insights into the research landscape, and optimised data curation efforts. SciLake's neuroscience pilot aims to enhance connectivity and FAIRness, benefiting researchers using EBRAINS data services and interoperable platforms globally.

About EBRAINS

EBRAINS is an open research infrastructure that gathers high-quality research data, tools and computing facilities for brain-related research, built with interoperability at the core. The infrastructure offers an extensive range of FAIR brain data sets, a most comprehensive multilevel brain atlas, AI-based tools for analysis, modelling and simulation, and access to high-performance computing resources, robotics and neuromorphic platforms to researchers. Explore the tools and services available [here](#).

EBRAINS AISBL is an international non-profit association, headquartered in Brussels, Belgium. It is organised around a central hub that coordinates a pan-European network of services delivered through currently 11 [National Nodes](#): Belgium, Denmark, France, Germany, Greece, Italy, Netherlands, Norway, Spain, Sweden, and Switzerland.

About SciLake

SciLake is a project funded by the European Union's Horizon Europe program. The project aims to create a seamless integration between domain knowledge and open Scientific Knowledge Graphs while also developing useful added-value services for specific research areas. The ultimate goal is to empower researchers and foster a more interconnected and efficient scientific community. SciLake brings together a competent consortium of 13 partners from 9 different countries. The consortium consists of partners with expertise in knowledge management and discovery, as well as experts from Neuroscience, Cancer, Transportation, and Energy research, who are involved in piloting activities.

Contact:

Archana Golla, PhD
Scientific data life cycle manager
EBRAINS Research Infrastructure
archana.golla@medisin.uio.no

###

8.2. Pilot Action Plans

This annex outlines the pilot action plans, including the key activities expected to take place in the context of each pilot during the creation and enrichment of the respective SKGs and testing or using the developed technologies in practice.

Pilot involved	Actions		Leading partners
Energy	E1	Extract the Energy-related subgraph of the OpenAIRE Graph (using EXPLORE)	HES-SO, OAIRE, CNR
	E2	Transform subgraph files and load them into an AvantGraph instance	TUe
	E3	Create a dedicated space with basic configuration in BIP! Spaces for Energy Research	ARC, HES-SO
	E4	Produce the subset of OA publications (collected by OAIRE using PDFfetcher) related to Energy Research	OAIRe, CNR, ICM
	E5	Use Automatic Translation tool to translate non-english titles & abstracts from selected languages (es, fr, pt)	ARC
	E6	Collect relevant legal documents from eurlex (or other source)	ARC, HES-SO
	E7	Combine NER with DoStRe to extract GeoAreas, CaseStudyDates & EnergyTypes from OA Publications	SIRIS/OPIX, DFKI, HES-SO
	E8	Use NER on legal documents to extract EnergyType-LegalDoc relationships	SIRIS/OPIX, HES-SO
	E9	Load EnergyType entities, GeoArea & CaseStudyDate fields in Publications, and EnergyType-Publication relationships to AvantGraph	TUe, SIRIS/OPIX
	E10	Load EnergyType-LegalDoc relationships that have been extracted to AvantGraph	TUe, SIRIS/OPIX
	E11	Use SciNoBo RAA on OA publications to identify Paper-Software & Paper-Dataset relationships	ARC
	E12	Use SciNoBo CA with DoStRe on OA publications to classify citations	ARC, DFKI
	E13	Load info from SciNoBo RAA and CA to the pilot's SKG in AvantGraph	TUe, ARC
	E14	Use sHINER and/or SciNeM to identify missing links in the SKG and include them to the SKG	TUe, ARC
	E15	Experiment with SKG contents (using AvantGraph queries, Lake API queries, BIP! Spaces UI, or/and SciNoBo assistant UI) -- this will conclude basic evaluation and experimentation	HES

Transport Maritime	M1	Extract the Maritime-related subgraph of the OpenAIRE Graph (using EXPLORE)	CERTH, OAIRE, CNR
	M2	Transform subgraph files and load them into an AvantGraph instance	TUe
	M3	Create a dedicated space with basic configuration in BIP! Spaces for Transport-Maritime Research	ARC, CERTH
	M4	Produce the subset of OA publications (collected using PDFfetcher) related to Transport-Maritime Research	OAIRE, CNR, ICM
	M5	Load VesselType entities/nodes (from AIS) to AvantGraph	TUe
	M6	Collect Regulation texts & metadata (from IMO exploiting R2PGM) & load Regulation entities & Regulation-Organisation relationships to AvantGraph	CERTH, ARC, TUe
	M7	Use Automatic Translation tool to translate non-english titles & abstracts from selected languages (es, fr, pt)	ARC
	M8	Use NER with DoStRe to extract mentions of VesselTypes from OA Publications coming from the Maritime-SKG	SIRIS/OPIX, CERTH, DFKI
	M9	Use NER to extract mentions of VesselTypes from Regulation texts identified	SIRIS/OPIX
	M10	Load VesselType-Publication relationships that have been extracted to AvantGraph	TUe, SIRIS/OPIX
	M11	Load VesselType-Regulation relationships that have been extracted to AvantGraph	TUe, SIRIS/OPIX
	M12	Use SciNoBo RAA on OA publications to identify Paper-Software & Paper-Dataset relationships	ARC
	M13	Use SciNoBo CA with DoStRe on OA publications to classify citations	ARC, DFKI
	M14	Load info from SciNoBo RAA and CA to the pilot's SKG in AvantGraph	TUe, ARC
	M15	Use sHINER and/or SciNeM to identify missing links in the SKG and include them to the SKG	TUe, ARC
	M16	Experiment with SKG contents (using AvantGraph queries, Lake API queries, BIP! Spaces UI, or/and SciNoBo assistant UI)	CERTH
Cancer	C1	Extract the Cancer-related subgraph of the OpenAIRE Graph (using EXPLORE)	CERTH, KI, OAIRE, CNR
	C2	Create a first version of the Cancer SKG using CKG and load in AvantGraph	CERTH, KI, ARC, TUe
	C3	Create a dedicated space with basic configuration in BIP! Spaces for Cancer Research	ARC, CERTH, KI

	C4	Incorporate content from the subgraph into the SKG	ARC
	C5	Update CKG with additional entities from other sources (maybe exploiting R2 PGM)	CERTH, KI, ARC, TUE
	C6	Produce the subset of OA publications (collected using PDFfetcher) related to Cancer Research	OAIRE, CNR, ICM
	C7	Use Automatic Translation tool to translate non-english titles & abstracts from selected languages (es, fr, pt)	ARC
	C8	Use NER and DoStRe to extract Genes from OA Publications	SIRIS, DFKI
	C9	Load Gene-Publication relationships in AvantGraph	TUE, SIRIS
	C10	Use SciNoBo RAA on OA publications to identify Paper-Software & Paper-Dataset relationships	ARC
	C11	Use SciNoBo CA with DoStRe on OA publications to classify citations	ARC, DFKI
	C12	Load info from SciNoBo RAA and CA to the pilot's SKG in AvantGraph	TUE, ARC
	C13	Use sHINER and/or SciNeM to identify missing links in the SKG and include them to the SKG	TUE, ARC
	C14	Experiment with SKG contents (using AvantGraph queries, Lake API queries, BIP! Spaces UI, or/and SciNoBo assistant UI)	CERTH, KI
Transport CCAM	T1	Extract the CCAM-related subgraph of the OpenAIRE Graph (using EXPLORE)	ICCS, OAIRE, CNR
	T2	Transform subgraph files and load them into an AvantGraph instance	TUE
	T3	Create a dedicated space with basic configuration in BIP! Spaces for Transport-CCAM Research	ARC, ICCS
	T4	Produce the subset of OA publications (collected using PDFfetcher) related to Transport-CCAM Research	OAIRE, CNR, ICM
	T5	Use Automatic Translation tool to translate non-english titles & abstracts from selected languages (es, fr, pt)	ARC
	T6	Use NER and DoStRe to extract domain-specific entities from OA Publications	SIRIS/OPIX, ICCS
	T7	Load extracted domain-specific entities in AvantGraph	TUE, SIRIS/OPIX
	T8	Use SciNoBo RAA on OA publications to identify Paper-Software & Paper-Dataset relationships	ARC
	T9	Use SciNoBo CA with DoStRe on OA publications to classify citations	ARC, DFKI
	T10	Load info from SciNoBo RAA and CA to the pilot's SKG in AvantGraph	TUE, ARC

	T11	Use sHINER and/or SciNeM to identify missing links in the SKG and include them to the SKG	TUe, ARC
	T12	Experiment with SKG contents (using AvantGraph queries, Lake API queries, BIP! Spaces UI, or/and SciNoBo assistant UI)	ICCS
Neuroscience	N1	Extract the Neuroscience-related subgraph of the OpenAIRE Graph (using EXPLORE)	UniO
	N2	Extract a subgraph of the EBRAINS KG and load it into an AvantGraph instance	UniO, TUe
	N3	Create a dedicated space with basic configuration in BIP! Spaces for Neuroscience	ARC, UniO
	N4	Produce the subset of OA publications (collected using PDFfetcher) related to Neuroscience Research	OAIRE, CNR, ICM
	N5	Use Automatic Translation tool to translate non-english titles & abstracts (from datasets) from selected languages (es, fr, pt)	ARC
	N6	Use NER and DoStRe to extract ExperimentalApproaches, Techniques, PreparationTypes, Parcellations, BiologicalSex & Species from OA Publications	SIRIS/OPIX, UniO
	N7	Load extracted ExperimentalApproaches, Techniques, PreparationTypes, Parcellations, BiologicalSex & Species entities in AvantGraph	TUe, SIRIS/OPIX
	N8	Use SciNoBo RAA on OA publications to identify Paper-Software & Paper-Dataset relationships	ARC
	N9	Use SciNoBo CA with DoStRe on OA publications to classify citations	ARC, DFKI
	N10	Load info from SciNoBo RAA and CA to the pilot's SKG in AvantGraph	TUe, ARC
	N11	Use sHINER and/or SciNeM to identify missing links in the SKG and include them to the SKG	TUe, ARC
	N12	Experiment with SKG contents (using AvantGraph queries, Lake API queries, BIP! Spaces UI, or/and SciNoBo assistant UI)	UniO