Operační program
Jan Amos Komenský

STUDIE PROVEDITELNOSTI
- VÝZVA OPEN SCIENCE II -

EXTRACT

KEY ACTIVITIES AND MAIN PLANNED OUTPUTS

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OPJAK.cz
MSMT.cz

Contents

10. 1. KA 2 – THEMATIC CLUSTER BIO/HEALTH/FOOD

PARTICIPATING PARTNERS: IOCB AS CR (TKA Guarantor), MU, UPOL

Biomedical and biochemical research in the Czech Republic currently suffers from data fragmentation and the absence of a common infrastructure. Significant data – clinical, image, simulation, or omic data – are often stored locally, unstructured, or are not shared at all. System support for standardization, long-term archiving and FAIR compatibility is absent. At the same time, however, in the Czech Republic there are top workplaces (e.g. BIOCEV, CEITEC) and infrastructure (e.g. ELIXIR CZ, Czech-BioImaging, BBMRI CZ), which generate world-class data and are actively involved in European projects.

The aim of KA 2 is to build repository infrastructure for the management, sharing and reuse of diverse types of biological, chemical and clinical data generated in both basic and applied research in the fields of medicine, biology, chemistry and related sciences. Repository sub-activities share a common framework that is firmly anchored in FAIR principles[1] – both in technical terms (standardized metadata, access interfaces, formats) and in procedural terms (curating, versioning, permission management, citation).

The individual repositories are designed as separate but interoperable components of the modular ecosystem. This creates an infrastructure in which complementary data about one subject (human, biomolecule, experiment) can be stored across multiple repositories and linked using metadata links, persistent identifiers and link integrations (e.g. via S3, RDF, Linked Data). The specific focus of each repository is as follows:

1) ClinData – human data, including clinical trials;

2) OmiCZ – multiomic data (e.g. genomic, transcriptomic, proteomic);

3) Imaging – medical and biological image data;

4) BioSimCZ – molecularly dynamic and structural simulation data;

5) Chemical biology (ChemBio) – chemical-biological data, machine learning tools (MS/AI);

An essential feature of all repository systems is to ensure interoperability with the National Repository Platform (NRP), the use of unified authentication/authorization through Life Science Login[2] and the possibility of exporting metadata to national and European catalogues and infrastructures (e.g. BBMRI-ERIC, ECRIN-ERIC, ELIXIR, EOSC, EBRAINS). FAIRification is not regarded as a one-off output, but as an ongoing, living framework that enables the effective and sustainable sharing and reuse of data across domains and institutions. For selected repositories, FAIRification is further extended to include direct support for advanced data processing tools (e.g. AI models in chemical biology, data annotation using ontologies).

This key activity thus does not represent an isolated effort, but a systematic change in access to data in the fields of health, biology and chemistry in the Czech Republic, with an emphasis on international compatibility, repeatability of research, openness and sustainability. The goal is to create

[1] https://faircookbook.elixir-europe.org/content/recipes/introduction/brief-FAIR-principles.html.
[2] https://lifescience-ri.eu/ls-login/.

an infrastructure that will allow not only the storage, but also the full use of data for advanced research (meta)analysis, education[3], inter-domain cooperation and knowledge transfer.

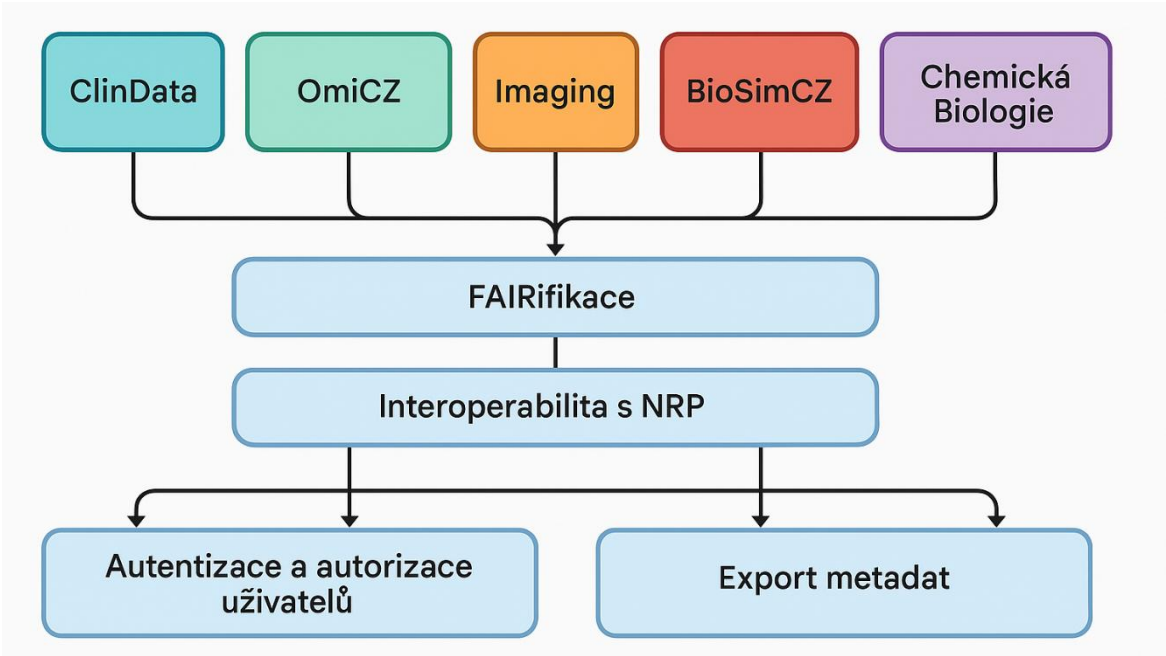

Fig. 3: Graphical diagram of activities

10.1.1. SUB-ACTIVITY 2.1 – NATIONAL OMICS REPOSITORY – CZECH OMICS NODE (OMICZ)[4]

The OmiCZ repository specializes in the storage, management and analysis of omic data (especially genomic, transcriptomic, proteomic and flow cytometric data), which are often sensitive and subject to the GDPR[5]. The goal of the repository is to provide secure, interoperable and FAIR-compatible solutions for working with these data at the national level, with full control over their availability and processing, in accordance with legal and ethical requirements and with connection to European FEGA and Beacon infrastructures.

Based on the analysis of the needs of scholarly communities, a metadata model compatible with European standards will be created (e.g. EGA/FEGA) using existing ontologies and controlled dictionaries.
Based on the metadata model defined in this way, the OmiCZ repository will be created as an instance of the basic NRP Invenio repository system, which will ensure efficient data storage, metadata management and secure access using the AAI unified authentication infrastructure.

As part of OmiCZ, software for the direct recording of omic data from devices (sequencers, flow cytometers) as well as user interfaces (ProducerUI, ClinUse, CropUI) will be implemented to enable

[3] Follow-up training activities to support data support staff and other experts within the community will be financed and organized by the EOSC CZ Training Centre, which will ensure adequate capacities and technical facilities.
[4] In the description of the KA, the planned activities are linked to the implemented outputs. The outputs are marked with a code. For TKA B/H/F, this is in the format **B_x** (key activity), B_x.x (partial, i.e. part of the key activity).
[5] https://mv.gov.cz/gdpr/clanek/co-je-gdpr.aspx.

domain-specific work with the repository, with the main added functionality of enabling the configuration and launch of analytical workflows.

The data and metadata stored in the repository will be interoperable not only at the national level (connection to clinical data repositories and NRPs), but also at the European level, thanks to integration with FEGA and Beacon network infrastructures.

An important part of the repository will be support for the secure processing of sensitive data through delegated computing workflows, where users will not be shown the primary data, but only the result of the workflow. The service will include an evaluation of the sensitivity of workflow outputs in accordance with ethical and legislative standards. It will also include the possibility of creating an interactive Galaxy environment within the Trusted Research Environment (TRE) to enable the secure processing of sensitive genomic data in the repository. The design and creation of these services to be used in the OmiCZ repository are planned and defined within the TKA SENSI.

Preparation of the OmiCZ system will also include a collection of test datasets and the design of real use scenarios that will cover various types of omic data (genomic, transcriptomic, proteomic, flow cytometric) as well as various application contexts (clinical data, plant/agro data, etc.). The datasets will be provided by the project partners and will be used to verify the functionality of individual system components, including recording, metadata management, access control and analysis in TRE.

In the final phase of the activity, training will be carried out for end users of the system and data support staff and other experts within the community (researchers, data curators, administrators). The goal of the training will be to ensure sufficient understanding of the management and use of sensitive data in the OmiCZ environment and the connection to the European infrastructure.[6]

ACTIVITIES:

- collection of requests from communities (types of data produced, data formats used, metadata requests);
- creation of a metadata model suitable for omics data that is compatible with EGA/FEGA, compliant with FAIR requirements and utilizes existing B_1.1 **ontologies**;
- implementation of the repository core on the Invenio platform, including authentication and authorization via AAI **B_1.2**;
- development and implementation of tools for data recording and management:
    - ProducerUI interface for data producers,
    - ClinUse for clinical data,
    - CropUI for plant/agro data,
    - deployment of a functional system enabling data recording and metadata management via ProducerUI and other interfaces. **B_1.3**
- ensuring of interoperability with European and national data space:
    - connection to FEGA and Beacon,
    - linking with ClinData via ClinLink,
    - export of metadata to a SPARQL engine (where relevant).
- functional interoperability with the European infrastructure (FEGA/Beacon) and ClinData, verified by exporting metadata and connecting to the SPARQL engine **B_1.4**;
- implementation of services for workflow and analytical tools:

[6] The training courses will be realized in cooperation with the EOSC CZ Training Centre, which will provide adequate capacities, technical facilities and organizational support and are not included in the budget for OS II.

- implementation of a Federated Workflow Execution service for batch analysis without data access and implementation of Galaxy service in TRE for interactive analysis **B_1.5**.

– preparation of test datasets and use scenarios:
  - datasets provided by partners covering different types of data and uses for system validation,
  - testing and preparation of documentation.
– governance framework (legal, ethical and operational aspects);
– user manuals and validation reports **B_1.6**.

SUB-ACTIVITY OUTPUT CODES[7]

B_1	National Omics Repository – Czech Omics Node (OmiCZ)
B_1.1	Architecture design documentation and OmiCZ repository metadata model
B_1.2	Functional prototype of the OmiCZ repository with integrated AAI
B_1.3	ProducerUI functional interface for data and metadata management
B_1.4	Linking the OmiCZ repository to FEGA, Beacon and ClinData
B_1.5	System for safe start of workflow and Galaxy in TRE
B_1.6	Governance and methodological documentation for the OmiCZ repository

10.1.2. SUB-ACTIVITY 2.2 – STRUCTURAL AND SIMULATION DATA REPOSITORY (BIOSIMCZ)

The main objective of the sub-activity is to create a Czech structural simulation repository, abbreviated as BioSimCZ. This is a new repository that will enable the storage, management and further processing of structural simulation data (i.e. simulation outputs related to 3D structures of biomacromolecules). Specifically, these will be data from atomic molecular dynamics, coarse grain simulations, predicted ensemble structures, etc., for which an adequate storage area does not currently exist either in the Czech Republic or abroad. The repository will also include other selected groups of structural data for which there is currently no repository (e.g. annotation to nucleic acid structures).

The repository will be tested using data produced by institutions involved in this sub-activity (MU, UPOL and IOCB AS CR) as well as cooperating institutions (Charles University, UCT). After the repository has been created, every laboratory in the Czech Republic will have the opportunity to upload its simulation data to it (biomolecule simulations are run by several institutes of the Academy of Sciences in addition to universities – e.g. IOCB, IBT, and BTU). Information about the repository created will be disseminated through the ELIXIR bioinformatics infrastructure as well as at meetings and conferences of the structural bioinformatics community. We disseminate information about new bioinformatics databases and tools in the same way. The repository will operate on the same principle as Protein Data Bank: data will be uploaded directly by users, thus obtaining a persistent data identifier that they can use, for example, in a publication.

For searching within BioSimCZ, identifiers common for biomacromolecular structures (PDB ID[8], UniProt ID[9], etc.) and used in standard bioinformatics databases Protein Data Bank, AlphaFold DB, UniProt,

[7] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.
[8] Protein Data Bank identification code – identifies an entry within the Protein Data Bank database (https://www.rcsb.org/).
[9] UniProt identification code – identifies a record within the UniProt database (https://www.uniprot.org/).

CATH managed e.g. ESFRI with the ELIXIR infrastructure are used[10]. 3D structures of molecules (in PDB[11] and mmCIF format[12]) will also be used for searches.

Links with the following projects will be ensured:

- BioExcel[13] MDDB[14]: The MDDB project is preparing a network of European nodes storing molecular dynamic data generated in a given country within individual countries. This distributed project requires the creation in every participating country of an MDDB national node dedicated to storing
  and managing molecular dynamic data in that country. The architecture, management and involvement of the national node in the MDDB are the responsibility of the employees of the node. BioSimCZ is planned to become the Czech national node for MDDB.

BioSimCZ will be included in the list of ELIXIR CZ data sources and will become part of the bio.tools database[15], which describes bioinformatics databases and tools. In addition, BioSimCZ will be connected to key databases of biomolecular structures managed by the ELIXIR infrastructure (e.g. PDBe-KB[16] or the 3D-Beacons Network[17]). Architecture, development and testing shall be integral components of the project. Following the creation of the repository, data from the sources where these data are currently located (ZENODO, Figshare, OSF, etc.) will be imported into it. Note: Therefore, no new datasets will be created. These data will be imported into the repository for the following reasons: They will be used to test the functionality and robustness of databases, allow for much better further data processing using computational tools and motivate users to upload their data to the repository. The BioSimCZ repository will be part of the NRP repository platform; specifically, it will be implemented using the Invenio system. The BioSimCZ metadata schema will be built over the NRP metadata schema, take over its items and add additional items that are specific to BioSimCZ).

The repository will use a metadata scheme for structural simulation data (e.g. molecular dynamics, coarse grain simulation) as well as a metadata scheme for other selected groups of structural data (e.g. specific types of residues such as non-canonical amino acids or nucleobases). The metadata schema for simulation structured data will be based on the data schemas of simulation tools (e.g. GROMACS, AMBER, etc.), generalized and expanded with metadata from users (project name, contact person, simulation description, etc.). The metadata scheme for other selected groups of structural data will be created on the basis of the structural alphabet available in the DNATCO database[18].

Furthermore, the repository will use a software tool for reading metadata from simulation data to enable the storage of simulations in the repository and for their conversion into a data format

[10] ELIXIR (https://elixir-europe.org/) is an ESFRI project focused on the development and maintenance of a European infrastructure for the life sciences.
[11] PDB is a file data format used within the Protein Data Bank database (https://www.rcsb.org/) and other databases of biomacromolecular structures.
[12] mmCIF (Macromolecular Crystallographic Information File) is a file data format used within the Protein Data Bank database (https://www.rcsb.org/)and other databases of biomacromolecular structures.
[13] BioExcel(https://bioexcel.eu) is the European Centre of Excellence for Research in Computational Biology.
[14] MDDB is the designation for the MDposit project (https://mmb.mddbr.eu/).
[15] bio.tools (https://bio.tools/)is a database of software tools and databases in the life sciences.
[16] PDBe-KB (Protein Data Bank in Europe – Knowledge Base, https://www.ebi.ac.uk/pdbe/pdbe-kb/)is a database summarizing the various properties of biomacromolecule structures.
[17] Varadi, M., Nair, S., Sillitoe, I., Tauriello, G., Anyango, S., Bienert, S., ... & Velankar, S. (2022). 3D-Beacons: decreasing the gap between protein sequences and structures through a federated network of protein structure data resources. *GigaScience*, *11*, giac118.
[18] DNATCO (https://dnatco.datmos.org/), published here: Černý J., Božíková P., Malý M., Tykač M., Biedermannová L., Schneider B. (2020). Structural alphabets for conformational analysis of nucleic acids available at dnatco.datmos.org. Biological Crystallography, 76 (9), 805–813.

according to the aforementioned metadata scheme for simulation structural data. This tool will also be used to check the internal consistency of simulation data and to convert other selected groups of structured data into a data format according to the aforementioned metadata scheme for these groups of structured data. This tool is an essential part of the repository because it guarantees the FAIRness of the data imported into it.

The BioSimCZ repository will be developed in two stages, first as a prototype and then as a complete version.

- The **prototype** will realize the storage, sharing and processing of data from molecular dynamics. An indispensable part of the prototype will be a tool for reading metadata from molecular dynamics data and converting them into a metadata schema, ensuring the FAIRification of the data and its import into the repository.
- With the incorporation of the prototype, the **complete version** of the repository will now allow the storing, sharing and processing of data from other types of simulations and work with other groups of structural data. A further essential component of the repository will be a tool for reading metadata from the aforementioned types of simulation data as well as from other groups of structural data and their conversion into the aforementioned metadata schemes. This tool, which ensures the FAIRification of data and their import into the repository, will also be able to check the consistency of simulation data.

In the final phase of the activity, training will be carried out for end users of the system, data support staff and other experts within the community (researchers, data curators, administrators).[19]

ACTIVITIES:

- design of the repository prototype architecture and a tool for the automatic reading of metadata;
- creation of a metadata schema for data from molecular dynamics **B_2.1**;
- implementation, configuration, deployment and testing of the repository prototype (including a tool for reading metadata, using the metadata schema created);
- creation of documentation for the repository prototype (including a tool for reading metadata and a metadata schema) **B_2.2**;
- design of the architecture of the complete version of the repository (including the architecture of the tool for the automatic reading of metadata);
- creation of a metadata schema for the aforementioned simulation data as well as a metadata schema for other types of structural data **B_2.3**;
- implementation of configuration, deployment and testing of the complete version of the repository (including a tool for reading metadata);
- creation of documentation for the complete version of the repository (including a tool for reading metadata and a metadata schema) **B_2.4**;
- Design and implementation of the connection of the complete version of the repository to international data sources (MDDB and ELIXIR sources) **B_2.5**.

SUB-ACTIVITY OUTPUT CODES[20]

[19] The training courses will be realized in cooperation with the EOSC CZ Training Centre, which will provide adequate capacities, technical facilities and organizational support and are not included in the budget for OS II.

[20] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

B_2	Structural Simulation Data Repository (BioSimCZ)
B_2.1	Documentation of the design of the repository prototype architecture (including the design of the architecture of the metadata reading tool and metadata schema)
B_2.2	Functional prototype of the BioSimCZ repository
B_2.3	Documentation of the design of the architecture of the complete version of the repository (including the design of the architecture of the tool for reading metadata and metadata schemas)
B_2.4	Complete version of the BioSimCZ repository
B_2.5	Functional interconnection of the complete version of the repository, connected to MDDB and ELIXIR data sources

10.1.3. SUB-ACTIVITY 2.3 – REPOSITORY FOR HUMAN AND ANIMAL IMAGE AND PHYSIOLOGICAL MULTIMODAL DATA (IMAGING REPOSITORY)

The repository for human and animal imaging and physiological multimodal data (Imaging Repository) will be used to safely store, manage and share data obtained mainly by magnetic resonance imaging (MRI), electroencephalography (EEG), positron emission tomography (PET) and computed tomography (CT). The repository will support the storage of data from research studies with human participants, as well as from animal studies.

The project focuses on the following key areas:

- **Technical implementation of the repository** in the NRP environment, specifically using the Invenio repository system. Available UI templates will be tailored to thematically focused metadata models. The designing of these metadata models will be based on an analysis of established ontologies and standards (e.g. OpenMINDS, bids) in order to reflect the specific aspects of individual data types (e.g. human vs. animal data, brain vs. cardiological, etc.). The resulting design will then be implemented into the repository system.
- **Ensuring the protection of sensitive data** through access control at the level of individual datasets using AAI components available in the NRP for the Invenio repository system in order to maximize the FAIR principles when working with sensitive data. To take account of varying degrees of sensitivity, different data disclosure modes will be available – from open access, through automated approval, to manual assessment by the dataset administrator. All non-anonymized human data will be processed on the basis of informed consent.
- **Connection to Trusted Research Environments (TRE)**, which will enable secure interactive and batch work with data stored in the repository. Typical sets of (containerized) tools will be prepared, expanding the available TRE and components that provide access to selected repository data in TRE based on the access rights of a specific user. This integration will provide support not only for researchers analysing data, but also for data curators in the preparation of data for publication and sharing.
- **Interconnection with the European research infrastructure** at the level of export of selected metadata records, which will ensure interoperability and the wider use of datasets in the repository. To cover the neuroscience community, we will utilize the EBRAINS infrastructure; specifically, tools for the preparation of datasets for the EBRAINS Knowledge Graph curatorial process, focusing on supporting datasets falling into the category of sensitive data, will be prepared.

- **Testing and validation of all components of the repository**, ensuring the suitability of functional aspects (user interface, tools, access rights), as well as the workflow of data management and publication. Particular emphasis will be placed on work with sensitive data in accordance with ethical and legal requirements.

ACTIVITIES:

- design of technical architecture, authorization, and metadata model;
- implementation of a repository allowing the development of downstream components and further development of metadata models;
- integration of a system for the authentication and authorization of access to data at the dataset level;
- design and implementation of repository user interfaces (UI for searching and accessing data; UI for dataset administrators; UI for data stewards and curators);
- development of tools for data stewards and curators, including tools for the annotation, validation and preparation of datasets, focusing on interoperability with the BIDS standard;
- linking to TRE for the purpose of working with sensitive data;
- software configuration of the TRE instance to suit the main target communities;
- ensuring interoperability with the European Digital Research Infrastructure;
- development of tools for exporting metadata and supporting the curatorial process for EBRAINS Knowledge Graph;
- testing and validation;
- preparation of system documentation and user manuals.

SUB-ACTIVITY OUTPUT CODES[21]

B_3	Repository for human and animal image and physiological multimodal data (Imaging repository)
B_3.1	Documentation of the technical architecture of the repository and of metadata models
B_3.2	Functional prototype of the repository core on the Invenio platform
B_3.3	Functional repository user interfaces for different types of users
B_3.4	Connecting the repository to TRE with user tools
B_3.5	Linking the repository to the EBRAINS Knowledge Graph
B_3.6	Functional and validated full version of the repository; documentation.

10.1.4. SUB-ACTIVITY 2.4 – REPOSITORY FOR CHEMICAL BIOLOGY DATA AND ITS CONNECTION WITH THE DEVELOPMENT AND TRAINING OF AI MODELS OF MS

The repository specializes in biological and chemical data.

A suitable data model will be proposed based on an analysis of which data and which data formats are used in the community. This data model should meet FAIR requirements and utilize existing and used ontologies as much as possible.

[21] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

A suitable architecture and basic infrastructure that allows data to be stored in the repository will be designed for the data model designed in this way. It is assumed that the tools will be able to unify the data into the proposed data model.

This system will be expanded with the possibility of exporting metadata, and where possible data, to a suitable RDF engine that will allow searching (meta)data using the SPARQL language. The main goal will be to enable federated queries that allow the stored (meta)data to be linked to other, already existing data sources based on Linked Data technology.

ACTIVITIES:

- Collection of requests from the community (which data they produce and which data formats they use).
- Creation of a data model suitable for the given data and that meets FAIR requirements. **B_4.1**
- Implementation of data recording and management tools. The tools must ensure the transferability of data into a suitable FAIR form. **B_4.2**
- Linking and retrieving of data. The stored data or their metadata must be exportable to the SPARQL engine, which will enable their querying and connectivity using federated queries with other (external) SPARQL engines. **B_4.3**
- Provision of access and workflow. Access will be controlled by linking to a common authentication and authorization infrastructure. **B_4.4**
- Preparation of test datasets. Test datasets will be based mainly on sample data provided by individual partners.

Linking of the chemical biology repository with the development and training of AI models of MS

In addition to the chemical biology data repository itself, the sub-activity will include the creation of an interconnected environment for the development, training and validation of machine learning (AI) models focused primarily on the analysis of mass spectra. The repository will also serve as a source of annotated training data as well as a target platform for the deployment of the resulting models, creating a closed cycle between data storage, analysis and the back validation of results.

A mechanism for continuous updating of models based on newly uploaded, curatorially verified data will be implemented. The system will support two levels of model cycle management:

- Community approach (crowdsourcing) – researchers will be able to upload their own datasets and contribute to training or testing of models.

- Curatorial mode – an expert team will supervise the quality of training data, set validation scenarios and issue "production" versions of models.

The resulting models will be versioned, documented and published within the same infrastructure framework as the repository itself. They will be accessible via API or web interface, both for research use and for integration into other repositories and tools.

The development will also include the creation of tools for the:

- validation of training and test datasets;
- monitoring of the quality of predictions and subsequent adaptation of models;

- automatic annotation and design of chemical structures from mass spectra using RDF and ontologies (e.g. ChEBI, PubChem);
- record of computational history and provenance of data and models, in accordance with FAIR principles.

This sub-activity links the repository of chemical biology data with the development of artificial intelligence tools aimed at the interpretation of mass spectra. The goal is not only to ensure secure and standardized data storage, but also to create an environment in which these data can be used for automatic annotation, the prediction of chemical structures, and the training of AI models. Particular emphasis is placed on openness and community involvement through established processes. The resulting models and tools will be provided as open services and linked to semantic standards (e.g. RDF, SPARQL) and the knowledge base, which will ensure interoperability in the European and global research area. This sub-activity will significantly contribute to accelerating and refining the analysis of chemical data, expanding open science and strengthening the use of AI in bioinformatics and chemical biology. **B_5**

ACTIVITIES:

- design and development of tools for integrating the training environment with the FAIR repository;
- creation of infrastructure for the management of annotated training datasets, including validation and curation processes;
- implementation of mechanisms for the versioning and continuous updating of models (e.g. retraining, performance monitoring);
- development of tools for the training and validation of models in two modes: community (crowdsourcing)
  and curatorial;
- making the resulting models available in the form of a service (e.g. API, user interface, RDF exports);
- development of tools for automatic data annotation and the design of chemical structures using ontologies (ChEBI, PubChem, etc.);
- preparation of supporting documentation, manuals, sample datasets and recommendations for contributors.

SUB-ACTIVITY OUTPUT CODES[22]

B_4	Chemical Biology Data Repository
B_4.1	Metadata Model
B_4.2	SPARQL endpoint
B_4.3	Test dataset
B_4.4	Documentation
B_5	Linking of the chemical biology repository with the development and training of AI models for MS

[22] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

10.1.5. SUB-ACTIVITY 2.5 - CREATION OF A NEW CLINDATA REPOSITORY AND FAIRIFICATION OF DATA

The Clinical Repository (ClinData) offers the secure and structured storage, management and processing of clinical data (human data), especially for clinical data that are part of basic and applied clinical research. The repository may also store, for example, multiomic, image or other types of data related to the clinical data for a specific subject. The repository provides the option of archiving closed studies and integration with multiomic and image repositories. It also allows data to be connected to other relevant repositories using software-created converters.

ClinData itself will become a repository system in the NRP.

It will facilitate the archiving of closed studies (adding repository functionality) and connection to multiomic and image data repositories, as well as to metadata directories and catalogues, such as LifeData[23][24], cBioPortal,[25]MOLGENIS, ECRIN-ERIC[26] and BBMRI-ERIC[27]. This will create a unique platform for the secure storage and integration of human data in a single location, and which will be widely used for research in medicine and biology, as well as the social sciences and humanities.

Implementation will focus, among other things, on the protection of sensitive data and connection to TRE[28].

The existing ClinData portal is designed for the continuous collection and management of clinical data. This system will be supplemented by the possibility of the permanent archiving of completed studies, including a dynamic metadata model and harmonization of the data model in OMOP[29].

The existing user authentication and authorization system will be supplemented with the possibility of the authentication of users, for example through LifeSci AAI. **B_6.1**

Sensitive clinical data will, to the extent specified by the user, be automatically exported to relevant domain metadata catalogues, where they will be analysed by future users. Pseudonymous or non-anonymous data will be made available to users on the basis of[30] a DTA concluded between the data owner and the user directly in the secure ClinData system. **B_6.2**

Human data often include genetic and other omics data as well as the results of imaging examinations (MRI, CT, PET/CT, PET/MRI, SPECT, ultrasound, etc., but also photographs, audio or audio recordings). In order to allow the effective analysis of these data at the individual level in relation to their clinical information (for example, laboratory examination, clinical picture of the disease, response to treatment, survival, etc.), we will link clinical, multiomic and image information related to the same individual in the ClinData repository in the form of links to an internal or external storage area. **B_6.3**

Ongoing testing of all repository functionalities. Preparation of supporting documentation, service board, user training and pilot deployment of the system for key research communities.[31] **B_6.4**

[23] LifeData portal: https://lifedata.cz/.
[24] https://www.cbioportal.org/.
[25] https://molgenis.org/.
[26] https://ecrin.org/, European Clinical Research Infrastructure Network.
[27] https://www.bbmri-eric.eu/.
[28] Trusted research environment.
[29] Observational medical outcomes partnership.
[30] Data transfer agreement.
[31] The training courses will be realized in cooperation with the EOSC CZ Training Centre, which will provide adequate capacities, technical facilities and organizational support and are not included in the budget for OS II.

ACTIVITIES

- design of architecture for the creation of a permanent repository for clinical data;
- interoperability of the authorization and authentication of ClinData users with LifeSci AAI **B_6.1**;
- linking to domain metadata directories and accessing of data **B_6.2**;
- linking to domain repositories **B_6.3**;
- testing **B_6.4**.

SUB-ACTIVITY OUTPUT CODES[32]

B_6	ClinData Repository
B_6.1	AAI
B_6.2	Metadata Model
B_6.3	Interoperability
B_6.4	Test dataset

10.1.6. SUB-ACTIVITY 2.6 – STANDARDIZATION AND SETTING OF PROCEDURES FOR STORING AND MANAGING ACCESS TO DATA

The current state of data storage and management from regulated clinical trials is characterized by a fragmented and unsystematic approach, where lack of standardization hinders the effective interoperability and identification of studies suitable for e.g. meta-analysis.

The expected target state after the completion of the activity includes the creation of uniform methodological procedures that will enable not only compliance with ECRIN standards[33] and increase the quality and availability of data, but also the integration of data between domain-specific repositories, for example between the omic repository and the clinical data repository. The implementation of a uniform structure for metadata and data will also significantly reduce the costs necessary for their harmonization and thus facilitate their sharing within the European infrastructure (e.g. GDI[34], FEGA, EOSC) (3.3). The adaptation of existing ECRIN metadata and data models will ensure full legal and ethical compliance with legislative standards (GDPR, IPR), thereby significantly reducing barriers to international and interdisciplinary research.

The activity will help to remove obstacles that prevent the FAIRification of data from regulated clinical trials, thus increasing the traceability and availability of such data to the broader professional community. The activity will give clinicians, researchers and other professionals a unified and interoperable system for the management of data from regulated clinical trials, facilitating access to data for meta-analyses and evidence-based medicine.

The two main thematic sections of the methodology – the creation of a standardized description of the source data (metadata) and the standardization of the data itself (structure, format) – will help create an interoperable system that meets the requirements of the ECRIN MetaData Repository (MDR) and the Data Sharing Repository (crDSR). This will enable a unified and interoperable data infrastructure that supports open science and efficient data sharing between European partners.

[32] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.
[33] ECRIN, the European Clinical Research Infrastructure.
[34] The Genomic Data Infrastructure (GDI).

A methodology for the FAIRification of data from completed regulated clinical trials will be created for medical data. The methodology builds on the clinical data repository as well as the omic repository, where it extends their use, interconnection and interoperability, especially in the context of the ECRIN research infrastructure.

ACTIVITIES:

- Research on the structure and format of general methodologies for the FAIRification of data and methodologies for the FAIRification of data from clinical research.
- Establishment of methodologies for standardizing, storing, and managing access to data from completed regulated clinical trials.
- Adaptation of the ECRIN metadata model for the methodology for the FAIRification of data from completed regulated clinical trials. **B_7**
- Collection of feedback on the methodology for the FAIRification of data from completed regulated clinical studies, its evaluation and implementation of changes.
- Creation of the final version of the methodology for the FAIRification of data from completed regulated clinical trials. **B_8**

SUB-ACTIVITY OUTPUT CODES[35]

B_7	Methodology for the FAIRification of data from completed regulated clinical trials
B_8	Adaptation of the ECRIN metadata model for the methodology for the FAIRification of data from completed regulated clinical trials.

10.1.7. SUB-ACTIVITY 2.7 – DEVELOPMENT AND PILOT IMPLEMENTATION OF TOOLS FOR THE DEVELOPMENT OF NDI FOCUSED ON INTEROPERABILITY AND USER COMFORT

Currently, we are facing multiple problems, for which the project includes targeted solutions. The current state is characterized by a number of problematic points: Fragmented research data with inconsistent formats. Limited interoperability between national and international repositories. Manual and time-consuming analysis of mass spectra. Insufficient use of AI for the automation of analytical processes. Low level of training for working with FAIR data and artificial intelligence.

Tools for the "Biological Imaging Data Repository" (**B_03**) are to be designed and developed. These are domain-specific tools that do not form part of the NRP due to thematically limited applicability. At the same time, these tools are necessary to ensure a wide portfolio of NDI services.

We will create standardized datasets and AI benchmarks for the validation and analysis of scientific data stored in the created repositories. Development of automated data management and annotation tools that minimize manual work. By linking national and European scientific infrastructure for easier data sharing.

Tools addressing a range of topics will be designed, developed and tested:

- prediction of chemical formulas using high-precision mass spectra and the integration of a standardized dataset and benchmark for artificial intelligence (Chemical Formula Prediction Tool);

[35] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

- creation of standardized workflows for the validation of data from mass spectrometers, focused on reproducibility and interoperability in the field of chemical biology (Validation and benchmarking of datasets);
- development of metadata standards and models ensuring interoperability between existing data repositories, e.g. ELIXIR CZ, EGA (FAIRification of datasets);
- automation of mass spectrometry data annotation using RDF and connection to open knowledge bases, e.g. ChEBI, PubChem (Semantic Data Annotation Tools).

Tools to increase user comfort when working with data from biological imaging **B_9**, in particular:

- annotation tools for the manual, semiautomatic and automatic annotation of data, including multimodal, multidimensional and time-lapse (development of tools for annotating image data from light and electron microscopy needed to create training and test data for machine learning algorithms, conversion between different annotation formats, checking the consistency of annotations, interconnection of developed tools with the data management platform and other NDI tools);

- implementation and integration of the Onedata platform for data management of microscopic images in light microscopy (ensuring the collection and transfer of data from smaller instrument PCs to larger server storage areas, data sharing between internal and external employees, extraction and viewing of instrument metadata from native files of the main manufacturers of light microscopes, assignment of persistent identifiers for public datasets, creation of clear documentation for users);

- implementation and integration of the Omero system with the data management platform and other NDI tools (deployment of the OMERO tool for easy viewing, analysis and annotation of image data, creation of tools for automated data and metadata imports, development of tools ensuring the functional interconnection of the OMERO system with the data management platform and other tools created within the project, automated linking of image and experimental data, SSO authentication and uniform authorization during communication between the Onedata tools and the Omero database);

- development of iRODS implementation for cryo-EM data lifecycle management (development and implementation of tools for the processing, management and publishing of data generated by electron microscopes, which differ significantly from the aforementioned light microscopes, using the federated iRODS cloud solution focusing on three key areas: (1) extension of the existing data transmission system to other types of scientific data and its implementation in other shared laboratories, (2) development of solutions for the automated deposition of cryo-EM data into domain-specific EMDB and EMPIAR repositories, including the integration of the "single-click" workflow for easy publication, (3) extension of real-time data analysis and automation of device settings using machine learning methods; the aim is to create a universal platform that will facilitate efficient data management throughout their life cycle, from acquisition through analysis to long-term storage and sharing in accordance with the principles of the European Open Science Cloud).

The services will be available and maintained with the support of the Czech-BioImaging infrastructure.

ACTIVITIES:

- design and development of tools, preparation of data models;
- introduction of the first standardized datasets and AI models;
- integration with EOSC and ELIXIR, testing with users;

- pilot implementation and testing;
- full implementation, open access to data models and workflow;
- completion of documentation, launch of services.

SUB-ACTIVITY OUTPUT CODES[36]

B_9	Key	Biological Imaging Data Repository Tools

10.2. KA 3 – THEMATIC CLUSTER OF MATERIAL SCIENCE AND TECHNOLOGY

PARTICIPATING PATRNERS: JH IPC AS CR (TKA Guarantor), VUT, UWB, MU, CU

Following the NRP project and PKA outputs, the key activity will create three environments for the quality management of FAIR research data in the field of material sciences and technologies in the Czech Republic. At the time of the project launch, these domains do not have their own field repositories, and semantic meanings and models are unstable and not enshrined at international level. The domain lacks tools for the direct storage of data in repositories, which would reduce the time needed for the high-level management, storage and reuse of data.

The central activity is the implementation of a new DANTE^c domain repository with an adequate metadata profile, selection of relevant licenses and a user interface that will help improve the storage, but also the searching and reuse, of research data. The emergence of a new repository represents a clear path to the consolidation of users from the domain cluster community. In addition to the general domain-specific data storage area, the repository and this KA will also focus on expanding the possibilities for automating individual steps of data management and FAIRification, including their extraction for the development of ML-AI approaches. For this purpose, highly specialized collections

with extended metadata profiles will be created in the repository for a detailed description of not only the stored data, but also the processes associated with the creation of these data and information about provenance. The semantic and structural settings of the repository and individual collections will be actively harmonized with similar activities worldwide in order to increase their interoperability. We expect that this will make the repository visible not only within the community itself, but also outside it, and improve the quality of stored FAIR data.

A need for storing outputs from complex research and technological, predominantly automated, processes has been identified in the domain cluster community. For this purpose, the DANTE^c repository will be adapted to work with hierarchical records. This will enable the manual, but mainly automated, storage of data from complex processes (workflows) directly into the repository, the interconnection of related data in the repository, and the recording of provenance information according to the latest knowledge in the field. The repository will be connected to data management tools for users and within research infrastructures and one tool for subsequent work with data in the repository. This interconnection will contribute to the improvement of stored domain-specific data and will also facilitate the broad inter-domain use of the repository using modern ML and AI-based tools.

[36] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

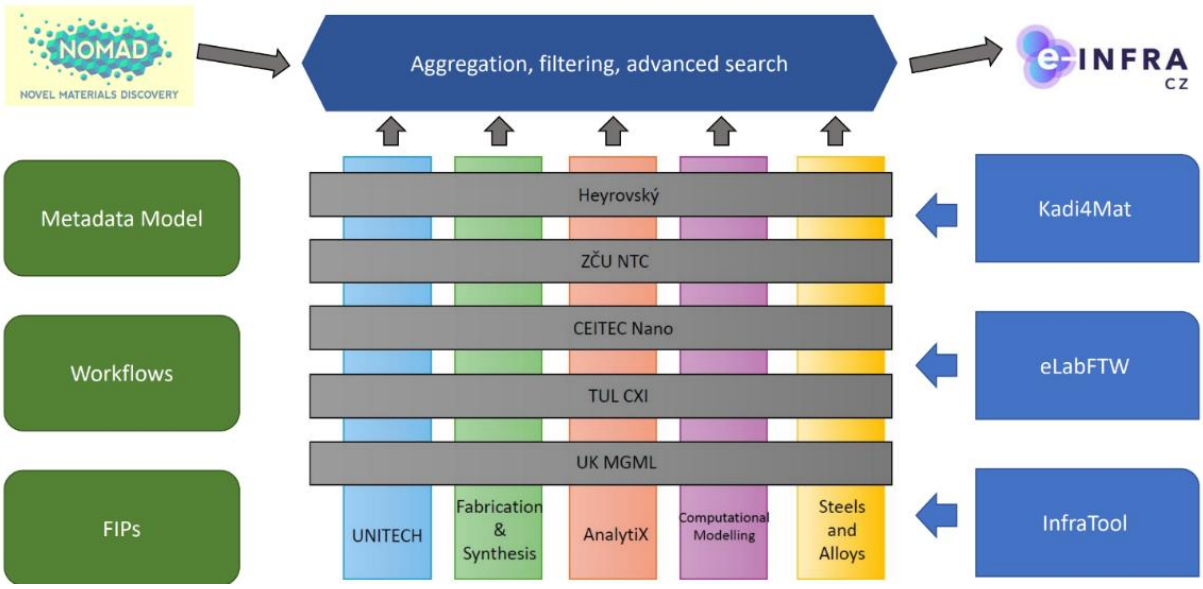

Fig. 4: Graphical diagram of activities

10.2.1. SUB-ACTIVITY 3.1 – CREATION OF A NEW DANTE^C REPOSITORY AND THE FAIRIFICATION OF DATA[37]

One new DANTE^c repository will be created with five collections (custom metadata, custom UI) for specific data. As part of the activity, five communities with their own curator and rules for the record approval process will also be consolidated (JH IPC, NanoENVI, CEITEC Nano, NTC-ZČU, CXI TUL). Outside the Heyrovský community, these communities belong to research infrastructures. Rules will be created for access to data and online documentation for all parts of the DANTE^c repository (including an interactive UI assistant and user-oriented widgets on the datamaterilized.eu community website).

For individual collections of the DANTE^c repository, metadata profiles (5 collections = 5 metadata profiles), community standards (FAIR Implementation Profiles, FIPs[38]) and license profiles for sharing open data, but also data for which it will be necessary to control access on an individual basis (sensitive data), will be created. The rules and procedures for working with sensitive data in the DANTE^c repository will be drawn up as part of TKA SENSI. Prior to implementation into the DANTE^c repository, the proposed standards and processes will be analysed and refined with the international community. Similarly, the possibility of sharing data with other domains and solutions tracking the origin and changes of records will be monitored – provenance (cf. the planned methodology). The quality of stored data will be ensured through training and hackathons.[39] Various approaches to the FAIRification of data and their storage in the repository will also be addressed here (e.g. using the repository superstructure for theoretical data or from the PKA area of the ELN).

As part of the project, we will create outputs that will improve the management and sharing of research data in the new repository. Benefits include: 1. five new metadata models and FIPs will improve

[37] In the description of the KA, the planned activities are linked to the implemented outputs. The outputs are marked with a code. For TKA MATECH, this is in the format **M_x** (key activity), M_x.x (partial, i.e. part of the key activity).
[38] https://www.go-fair.org/how-to-go-fair/fair-implementation-profile/.
[39] The training courses will be realized by the EOSC CZ Training Centre, which will provide adequate capacities, technical facilities and organizational support.

the FAIRificability of stored data[40]; 2. one new tool will be created to search, filter and aggregate related data for further use, e.g. for ML and AI training. Without this application, data preparation would be time-consuming and have no connection to the NOMAD data repository[41] (developed by the FAIRmat consortium, NFDI, Germany). The tool will increase data re-use[42]; 3. the NeXuS[43] parser and methodology for AI/ML will shorten the data processing time to international machine-processable standards and increase the possibility of data reuse using AI/ML; 4. outputs will improve the quality of work in a number of institutions.[44]

Given the overall scope of activities, we divide the description of activities into individual logical units below.

1/ New DANTE^{c} domain repository

The repository will utilize the Invenio CESNET software system (from the NRP project). This will simplify the connection of the DANTEc repository to AAI eINFRA, metadata mining to the NMD and registration in NCR. The creation of the repository also encompasses the:

i. **Creation of policies** for inserting records into collections in the DANTE^{c} repository and subsequent approval and editing/curatorial processes.

ii. **Creation of user interfaces** for the insertion of records into collections in the DANTE^{c} repository using the REACT/SB system (including design alignment with NRP/NDI).

iii. **Establishment of license profiles** for records in the DANTE^{c} repository using KA4.2 outputs of the NRP project.

iv. **Administration and governance** of the DANTE^{c} repository, including all five collections.

v. **Editing and curation of data** inserted into collections in the DANTE^{c} repository.

vi. **Creation of online documentation,** an assistant for UI repositories and user-oriented widgets on the community website (datamaterialized.eu).

At the end of the project, the DANTE^{c} repository will operate five collections with their own metadata profiles. The collections, i.e. sections for storing various types of data, will include: *UNITECH* – general section of the repository without strict specification of technical metadata, *Fabrication&Synthesis* – preparation of materials and synthesis of chemical substances, *AnalytiX* – characterization of materials and analytical procedures, *Computational Modelling – in silico* modelling of materials and their properties, and *Steels and Alloys* – database of steels and alloys. These collections were identified within the community at in-person MATECH EOSC WG CZ meetings in 2024. Individual collections will have different metadata profiles, and thus a different user interface for the adding of records by humans, or a different schema in the API for automated records from devices and ELN. The UNITECH collection will utilize the basic domain-specific metadata model DANTE^{core}, which will also be shared with other collections of the DANTE^{c} repository for greater interoperability. Within the community, we decided to take advantage of the fact that international repositories and data infrastructures for MATECH domains do not yet exist. In this, this community differs from other TKAs (e.g. BIO or HUMAN), which connect to the existing international system of data infrastructures and repositories within the NRP. This allowed us to unify the needs of the community into the single instance of the CESNET Invenio

[40] We assume 40–50 records with a FAIR score in the F-UJI application above 59%.
[41] https://nomad-lab.eu/nomad-lab/; https://www.fairmat-nfdi.eu/fairmat.
[42] We expect 20 re-used records by the end of 2028.
[43] https://www.nexusformat.org/.
[44] We expect at least 20 curated records by the end of 2026, with an increase of at least 100% in each subsequent year.

repository. The collections will then function as virtually separate repositories for the specific needs of the sub-domains. At the same time, it will be easier to manage the Invenio instance, connected SW tools and services and, we assume, data interoperability[45].

During the project, the DANTE^{c} repository will utilize at least five communities with their own data curator and workflows for the addition of records to the repository[46]. In the first phase, we expect data curators to form around collections and at the same time serve the entire domain cluster community. Following the implementation of good practices (expected at the end of the project), data curators will primarily serve their communities. Due to the process of gradually connecting the repository to automation tools, the communities will mainly use similar workflows/policies for the addition of records and their approval following publication during the project. We expect that this may change towards the end of the project due to the specific needs of the participating communities. The repository will be gradually connected to relevant services in the NRP according to their availability in the NRP environment.

The basic version of the DANTE^{c} repository connected to NRP services will use eINFRA AAI, enable the allocation of basic license selection and work with one metadata model (DANTE^{core}). During the project, other services that have been created or are in development under the NRP project will be connected to the repository. These are, in particular, the FAIRification tool[47], the automated entry of records using simplified tools,[48] ELNs created as part of the PKA of this project, and the selection tool for the selection of the licence[49]. This will result in a repository that is connected to the NRP, including selected advanced services and functions for working with a broad range of metadata models, including hierarchical records from complex instrumentation systems and research and development processes. The new DANTE^{c} repository will include detailed documentation in the form of an assistant in the user interface and an online document in the NRP document repository. The datamaterialized.eu community website will be supplemented with simplified instructions and aids to help users quickly master work in the DANTE^{c} repository. **M_1**

ACTIVITIES:

- DANTE^{c} repository connected to NRP services – basic version with connection to the NMD, eINFRA AAI and basic list of licences;
- implementation of communities and involvement of community administrators and curators i n the DANTE^{c} repository;
- repository linked to NRP services – production version;
- full functionality of the repository, including automatic data and metadata collection. **M_1**

2/ Metadata profiles and standards for the DANTE^{c} repository

For the DANTE^{c} repository, the basic NMD metadata profile (Czech Core Metadata Model; CCMM) will be primarily extended using the Metadata4Ing ontology[50] and the Korean data model

[45] Interoperability is ensured by using the same metadata model, thanks to which they will have the same basis and thus be internally interoperable.
[46] The communities are similar to the European Zenodo repository. We expect communities to set up institutions or subunits of institutions. Communities have their own "submission workflow" and their own curators.
[47] output of KA 5.1 of the NRP project.
[48] output of KA 5.3 of the NRP project.
[49] output of KA 4.2 of the NRP project.
[50] Ontology developed by the NFDI4Ing consortium in Germany , https://nfdi4ing.pages.rwth-aachen.de/metadata4ing/metadata4ing/; https://nfdi4ing.de/.

for the description of materials[51]. This will lead to the creation of the DANTE^{core} metadata profile with ontology for general-purpose, domain-specific data. This model will be used in *UNITECH* collection and form the basis for models of other collections in the DANTE^{c} repository. Efforts will also be made to harmonize this model with other domain-specific repositories in the NRP. For this purpose, we will analyse the metadata profiles used in other TKAS (e.g. TKA PHYSICS or TKA DM4AI). In the specialized collections (see above), we will focus on the technical and process parameters associated with the given sub-areas of research and the data flows associated with them. We will primarily use curated EMMO[52] and PMDco[53] ontologies and the general terminology defined in Schema.org[54]. Specific classes and attributes will then allow a detailed record of research processes in the aforementioned specialized collections. Only detailed records with rich information not only about the data itself, but also about the related processes, are usable for the creation of high-level ML and AI models. In order to verify the functionality and interoperability of the DANTE^{c} repository, we will perform a FAIR data analysis that will be continuously stored in the repository. We will use any shortcomings to modify the repository or the tools connected to it. **M_1.1**

The metadata profiles will be supplemented with data standards defined in the FAIR Implementation Profiles (FIPs), which we will create for the fields of material sciences and technologies using the FIP Wizard tool[55] for various types of use of the DANTE^{c} repository. The FIPs will define the standards supported, and in some cases even required, by the given collection and will also be used for the FAIR-ification of data. We expect to create multiple versions of the FIPs, even if a collection has not yet been created for the relevant specification in the DANTE^{c} repository. **M_2**

In the course of the project we will analyse (meta)data standards used or developed in similar activities abroad. Typical examples of these are activities in Germany (NFDI4Ing, NFDI-Matwerk[56], NFDI4Chem[57], Platform Material Digital – PMD[58]), France (DIADEM[59]), Denmark (CAPeX[60]), Norway (SINTEF[61]), Japan (DICE[62]), Korea (NCMRD) and the USA (MGI[63]). This analysis will also include the monitoring and cataloguing of technical interfaces (APIs) used by software systems developed within these activities. The goal of this process is the maximum interoperability of solutions designed for the DANTE^{c} repository. Detailed consultations on these analyses and work on the FIPs will then be held at the international level, specifically through RDA working groups (a group "Harmonised terminologies and schemas for FAIR data in materials science and related domains"[64] currently exists) or through direct negotiations with the coordinators of the aforementioned activities. This activity entails business trips abroad for the purpose of the harmonization of standards (metadata profile, ontologies, controlled dictionaries, file formats) through online consultations, as well as active participation in various in-person

[51] author Kwang-Ryeol Lee, National Centre for Materials Research and Development (NCMRD), Seoul, South Korea; https://github.com/krlee227/MatResData-Standard-Committee.
[52] https://emmo-repo.github.io/.
[53] https://materialdigital.github.io/core-ontology/.
[54] https://schema.org/docs/schemas.html.
[55] Output of KA 5.1 of the NRP project.
[56] https://nfdi-matwerk.de/.
[57] https://www.nfdi4chem.de/.
[58] Platform Material Digital, https://www.materialdigital.de/.
[59] https://www.pepr-diadem.fr/le-pepr/.
[60] https://capex.dtu.dk/.
[61] https://www.sintef.no/en/main-research-areas/materials/.
[62] https://dice.nims.go.jp/; National Institute of Materials Sciences, Tsukuba, Japan – https://www.nims.go.jp/eng/.
[63] Materials Genome Initiative, https://www.mgi.gov/;
[64] https://www.rd-alliance.org/groups/harmonised-terminologies-and-schemas-fair-data-materials-science-and-related-domains-wg/members/all-members/.

conferences and workshops focused on advanced materials for industry and their digitization, FAIR data, and open data. **M_1.1**, **M_1.2**, **M_2**

We will analyse the needs for integration of the DANTE^{c} repository with the current standards for storing provenance information. The output of this activity will be the definition of needs related to the creation of data at various levels and at various institutions using the provenance of metadata, in particular for the purpose of traceability of the origin/predecessors of datasets, both by the user requesting access to the data set and by the repository administrator or data curator. This activity will utilize the data from the Provenance PKA sub-activity to ensure that information on provenance is compliant with the international ISO 23494 standard, the underlying CPM data model, and will contribute to the reuse (R in FAIR) of data stored in the DANTE^{c} repository. **M_1.3**

ACTIVITIES:

- definition of (meta)data standards – FIP; **M_2**
- creation of the DANTE^{core} metadata model;
- extension of the metadata profile for individual collections, 5 collections = 5 profiles; **M_1.1**
- analysis of the compatibility of semantic standards and SW interfaces in an international environment; **M_1.2**
- analysis of the needs for integration of the DANTE^{c} repository with current standards for provenance; **M_1.3**
- analysis of the data call and hackathon.

3/ **SW tool for the synchronous search and filtering of data, including their extraction, in the DANTE^{c} and NOMAD repositories**

To make better use of the data stored in the DANTE^{c} repository, we will create a SW tool for the synchronous searching and filtering of data, including their extraction, in the DANTE^{c} and NOMAD repositories[65]. NOMAD is currently the most well-developed data repository for the material sciences in Europe. The interconnection of the DANTE^{c} and NOMAD repositories will take place in several stages. First, we will map the DANTE^{core} and NOMAD data models. If necessary, an extended DANTE^{core+} metadata model will be designed to facilitate the improved interoperability of both services. We will develop the SW tool to communicate with the APIs of both repositories and offer a unified means of filtering data in repositories and their subsequent acquisition for later processing (primarily for the development of AI tools in the material sciences). We will also set the functionality of a filtering and data processing tool for working with hierarchical records from automated processes. Finally, detailed instructions and a methodology of work will be prepared for this SW tool (in conjunction with KA7 NRP). **M_1.4**

ACTIVITIES:

- mapping of the DANTE^{core} and NOMAD metadata profiles,
- SW tool for the synchronization of work with data in the DANTE^{c} and NOMAD repositories. **M_1.4**

4/ SW tool Invenio NeXuS "parser" and methodology for AI/ML

We will develop the Invenio NeXuS "parser" tool for theoretical data and experimental data. This is plug-in for Invenia, which processes datasets uploaded to the DANTE^{c} repository and automatically

[65] I.e. a separate service with its own GUI supplementing the current methods of searching and filtering and aggregation.

provides technical metadata. The resulting "parser" reads the output formats of theoretical computing packages or the NeXuS format, the input output data of theoretical models and other data formats and converts the data into the expanded metadata format of the DANTE^c repository. Compliance with the NOMAD portal will also be checked at this point. A methodology will be developed for the FAIRification of data intended for the complex analysis of experimental and theoretical material data (primarily for AI/ML technologies) in the DANTE^c repository. **M_1.5**, **M_1.6**

ACTIVITIES:

- SW tool: Invenio NeXuS "parser" for theoretical and experimental data; **M_1.5**
- methodology for the FAIRification of data for AI/ML in the DANTE^c repository. **M_1.6**

SUB-ACTIVITY OUTPUT CODES[66]

M_1	DANTE^c repository
M_1.1	Metadata profiles for DANTE^c repository collections
M_1.2	Analysis of the compatibility of semantic standards and SW in an international environment
M_1.3	Analysis of provenance needs for the DANTE^c repository with current provenance standards
M_1.4	SW tool enabling synchronous work with data in the DANTE^c and NOMAD repositories
M_1.5	Invenio NeXuS "parser" SW tool for theoretical data and experimental data.
M_1.6	methodology for the FAIRification of data for AI/ML in the DANTE^c repository.
M_2	FAIR Implementation Profiles for communities in the DANTE^c repository

10.2.2. SUB-ACTIVITY 3.2 – DEVELOPMENT AND CONNECTION OF TOOLS AND SERVICES FOR THE COLLECTION OF DATA FROM RESEARCHERS AND INFRASTRUCTURES TO THE DANTE^C REPOSITORY

Both direct automation tools for the collection and processing of (meta)data (e.g. from scientific instruments or in combination with computing capacities)[67] and tools for the everyday work of scientists with their data are required in order to improve the environment in which research data is managed. These tools include electronic laboratory (field) notebooks, known as ELNs. PKA ELN deals with the integration of two internationally established ELN systems (Kadi4Mat[68] and eLabFTW[69]) in the NDI environment.

The aim of this activity is the assisted and semi-automated collection of records into the DANTE^c repository. We connect the repository to the two existing software tools for user data management (ELN) described above. The tools will simplify daily work with FAIR data, helping to improve the quality of data collected without unnecessarily increasing the burden on VO scientists and support staff. With these tools, the FAIRification of data moves closer to the moment of their creation and facilitates the

[66] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.
[67] Scope of the NRP project (KA 5.3 and KA 5.4).
[68] https://kadi.iam.kit.edu/.
[69] https://www.elabftw.net/.

continuous recording of provenance data. The connection of tools to the DANTE^c repository will further expand the options for working with one of the basic NRP repository systems – CESNET Invenio.

As part of this activity, we will also develop a new open-source SW package that will meet the requirements of infrastructure managers, i.e. a solution that enables the integrated management of users, devices and produced scientific data of the research infrastructure within a single software system. We will connect this software system to the DANTE^c repository for the direct collection of records from research infrastructures. Subsequently, this solution will be implemented in two research infrastructures focused on material research and technology development: MGML at the Charles University Faculty of Mathematics and Physics (CU MATHPHYS) in Prague[70] and the CzechNanoLab infrastructure laboratory[71] (at the Institute of Physics of the Czech Academy of Sciences (AS CR) in Prague and CEITEC Nano at the University of technology in Brno[72)].

This activity will utilize the results of the PKA ELN area, which integrates the software tools for the Kadi4Mat and eLabFTW ELNs into the NRP environment. As part of this activity, these ELNs will be directly connected to the DANTE^c repository and the metadata profiles will be reconciled.

As part of the activity, we implement the Kadi4Mat ELN in at least five research groups and prepare user documentation for direct deposition of records into the DANTE^c repository. The service integrating the Kadi4Mat ELN with the repository will allow the direct selection of data in the ELN and the sending of the data to selected collections of the DANTE^c repository. The service will also facilitate checking of the mandatory and recommended metadata for a given collection. In the event of non-compliance with the conditions for recording, the sender will be informed of the non-compliance or a form will be issued for the supplementing of the missing metadata. The service will also be adapted to work with hierarchical records from automated processes. Templates (sample records) for the ELN will be created, which will be consistent with the profiles of the collection of the DANTE^c repository. The repository for receiving data from the Kadi4Mat ELN instances will be modified. No general procedure is yet in place in the Czech Republic for interconnecting the ELN with the selected repository for the fields of material sciences and technologies. **M_3**

We will also implement the eLabFTW ELN in at least five research groups and prepare user documentation for the direct deposition of records into the repository. The service and templates will be developed, and the DANTE^c repository modified, in a similar way to that described for the Kadi4Mat ELN. **M_4**

We will develop an open-source SW tool (package) for the management of users, devices and their data outputs for large research infrastructures. The tool will allow users to register in the infrastructure, check their accounts, set access rights to devices and data repositories, book scientific devices (booking), automatically collect (meta)data, and check mandatory metadata for predefined storage areas. This tool will also ensure the immutability of the resulting data (transparency). The package includes the registration of devices and their technical parameters. These parameters will be automatically attached to the measured data to increase the reproducibility of the measurement. The tool will enable the administrative management of scientific instruments, including their maintenance. **M_5**

We will connect this software tool to the DANTE^c repository for assisted collection of records from research infrastructures. This service will allow infrastructure users to directly store the acquired data records and their metadata (including detailed technical parameters) in selected collections

[70] Materials Growth and Measurement Laboratory, https://mgml.eu/.
[71] https://www.czechnanolab.cz/.
[72] https://nano.ceitec.cz/.

of the DANTE^{c} repository. The tool will also be adapted to work with hierarchical records from automated processes. The repository for receiving data from the SW data management tool in research infrastructures will be modified. We will create user documentation for the created SW package for implementation in research infrastructures. Based on the performed user tests (including hackathons), we will then complete the methodology for the use of SW tools and sample examples of use. We implement this software tool in two Czech infrastructures (CzechNanoLab and MGML). **M_6**

ACTIVITIES:

- connection of the Kadi4Mat ELN to the DANTE^{c} repository; **M_3**
- implementation of the Kadi4Mat ELN in the research environment of material sciences and technologies;
- connection of the eLabFTW ELN to the DANTE^{c} repository; **M_4**
- implementation of the eLabFTW ELN in the research environment of material sciences and technologies;
- a comprehensive SW tool for managing users, devices, and their data outputs in research infrastructures, including methodology; **M_5**
- integration of SW tools for managing users, devices and their data outputs with the DANTE^{c} repository;
- implementation of a software tool for managing users, devices and their data outputs in the MgML and CzechNanoLab research infrastructures. **M_6**

SUB-ACTIVITY OUTPUT CODES[73]

M_3	Integration of the Kadi4Mat ELN and the DANTE^{c} repository
M_4	Integration of the eLabFTW ELN and the DANTE^{c} repository
M_5	SW tool for managing users and devices and their data outputs in research infrastructures
M_6	integration of SW tools for managing users, devices and their data outputs in research infrastructures with the DANTE^{c} repository

10.3. KA 4 – THEMATIC DATA CLUSTER MANAGEMENT FOR ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

PARTICIPATING PARTNERS: VSB (TKA Guarantor), MU, CU

There is no centralized platform for the sharing and storage of AI/ML data/models/workflow in the Czech Republic; frequently, the only available platforms are independent initiatives and projects, which often use platforms such as GitHub or Hugging Face to share their models and data, thus increasing their dependence on foreign platforms. The aim of this activity is therefore to establish a platform/repository along the same lines and that is embedded in the Czech environment, including the Czech and European legal environment with regard to the use of AI for sensitive data. This will significantly increase control over data, outputs and technology development in the field of AI/ML, consolidate the user community and strengthen independence from foreign infrastructures.

The aim of the creation of a new, domain-focused AI/ML (Artificial Intelligence/Machine Learning) repository Data Management for Artificial Intelligence and Machine Learning (DM4AI) based

[73] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

on the Clarin DSpace repository system is to offer a unified platform inspired by the global Hugging Face application. This platform will enable the effective sharing and management of AI/ML models, datasets, and workflows, including the provision of advanced tools for working with data and the possibility of connecting to computing infrastructures through the LEXIS Platform.

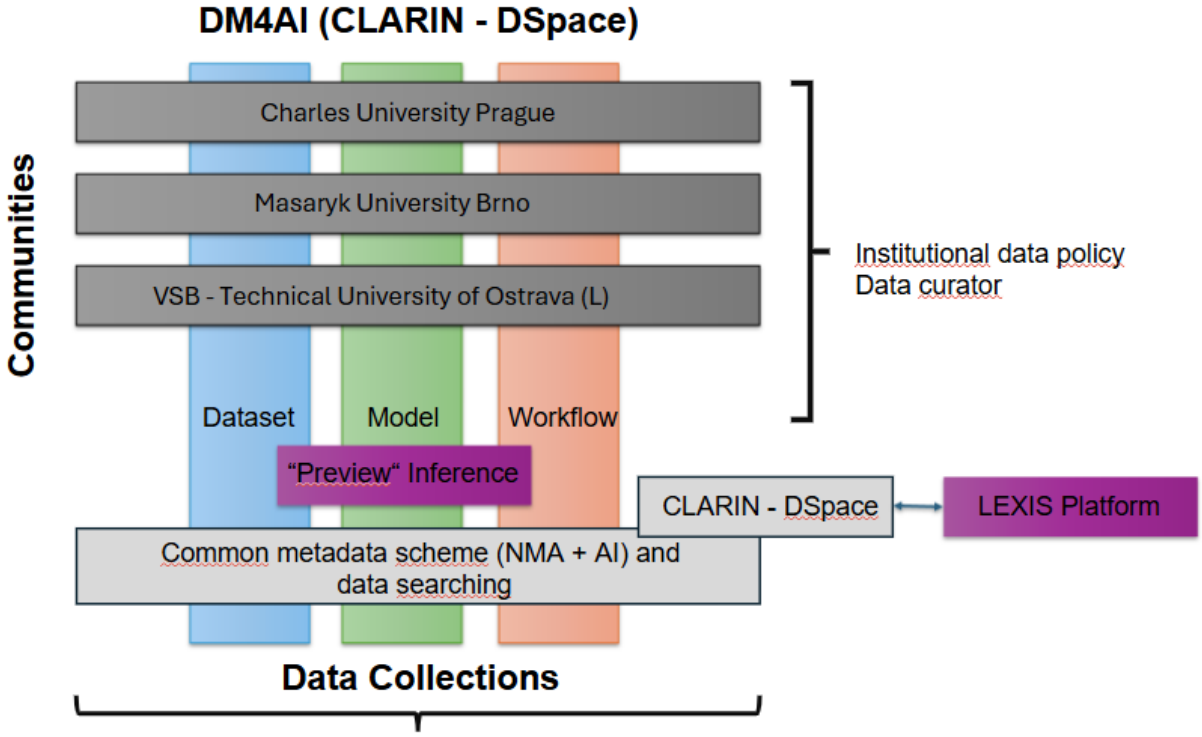

Fig. 5: Activity schema

The new repository platform (repository + all its tools) will reflect AI/ML domain specifics and will be fully adapted to them in terms of both structure and functionality in order to best support work with data, models, and AI/ML workflows. In terms of AI/ML specifics, this new repository will deliver several key features for researchers and the DM4AI community:

- Robust, complex and structured metadata that takes into account the specific attributes of AI/ML data, models or licenses.
- Separation of key parts of the AI/ML workflow (data, models) increases the clarity and reusability of data.
- The data in the repository will be managed in accordance with FAIR principles, ensuring their maximum quality and usability. In the field of artificial intelligence, data quality is of crucial importance – a model is only as good as the data on which it learns (e.g. the problem of reducing data bias).
- Artificial intelligence is heavily dependent on computing resources, as complex models and algorithms require a huge amount of computing capacity for effective training and inference, and therefore the AI/ML repository will provide the ability to connect to computing infrastructures while providing the function of simplified preview of AI/ML models directly within the repository environment ("preview" inference).
- The AI/ML repository will allow better control over data and their origin/transformations, licences, and compliance with local/global AI regulations (e.g. the AI Act[74]) and the use of AI for innovation.

[74] AI Act | Shaping Europe's digital future.

Members of the professional team involved in the project will undertake business trips abroad in order to acquire new knowledge by participating in foreign conferences/seminars/workshops, establishing contacts with foreign experts and disseminating the results of the project. In particular, these are trips related to participation in the European EOSC – participation in working groups (EOSC task forces), EOSC Symposium, EOSC winter/summer schools and related outreach activities, as well as e.g. Open Repositories, COLING and ACL conferences.

10.3.1. SUB-ACTIVITY 4.1 – CREATION OF DM4AI REPOSITORY INFRASTRUCTURE SUPPORTING FAIR PRINCIPLES AND INTERDISCIPLINARY INTEROPERABILITY[75]

The aim of the activity is to create a domain-specific repository for the domain cluster Data Management for Artificial Intelligence and Machine Learning that provides access to datasets, their metadata, AI/ML models, and their workflow, including the possibility of connecting to computing infrastructures. An instance of the Clarin-DSpace software repository system will be deployed.

The repository will be intended for datasets (existing or new) created by researchers with a link to AI/ML and will also include datasets (existing or new) created by other entities with which Czech researchers work and which are not available in other repositories. We will provide users with the option to store information about the entire AI/ML workflow that describes the input data used (available in this repository, or in the form of a link to an external data source), applications for AI/ML training (the algorithm and cluster used to run it, including the configuration of the environment), and the resulting model (e.g. neural network weights), which together constitute the FAIR Digital Object. The repository will facilitate efficient data management and search capabilities.

Given the overall scope of activities, we divide the description of activities into individual logical units below.

1/ New DM4AI repository, development of competences, and dissemination of the necessary know-how to domain-oriented research communities

The NRP infrastructure, LEXIS platform, Clarin-DSpace software platform and powerful e-Infra computing resources (e.g. Karolina computing cluster) will be utilized in the creation of the repository. The repository will primarily utilize the EOSC/e-INFRA CZ Common Authentication and Authorization Infrastructure (AAI) and will be properly registered in the NQF.

We will take full advantage of Clarin-DSpace, which was developed within the NRP with the creation of a specific metadata model for AI/ML models. The graphical user interface of the repository will be designed so as to meet the above requirements and will be inspired, among others, by the user interface provided by the popular Hugging Face service[76] (specifically: Model Hub[77], Datasets Library[78], Train[79] on external computing resources, taking into account the specifics of the AI and ML community). The interface of the deployed Clarin-DSpace instance will be adapted to meet the needs of the domain group, including the creation of the necessary metadata schemas and integration

[75] In the description of the KA, the planned activities are linked to the implemented outputs. The outputs are marked with a code. For TKA DM4AI, this is in the format **AI_x** (key activity), AI_x.x (partial, i.e. part of the key activity).
[76] https://huggingface.co. Platforms such as Kaggle or Hugging Face providing similar functionality, connected to commercial computing resource providers. These platforms allow the user to simplify direct interaction with the selected model for the first short assessment.
[77] https://huggingface.co/docs/hub/en/models-the-hub.
[78] https://huggingface.co/docs/datasets/en/index.
[79] https://huggingface.co/docs/transformers/en/training.

with the LEXIS platform[80] for distributed data management and the running of complex computing tasks on powerful clusters. **AI_1**

An integral component of this key activity is the creation of tutorials and specialized materials intended for new infrastructure users from the AI/ML community. These will be created across key activities in the domain cluster, and in particular will be linked to activities related to the creation and use of a domain-specific repository and software (SW). The main objectives of the output include: (1) Preparation of analytical materials related to the needs of research data management in relation to TKA Data Management for Artificial Intelligence and Machine Learning. (2) Preparation of "how to do" instructional materials (repository, software)/manuals/educational materials for users/lectures/methodologies (3) User training.[81] **AI_1.11**

ACTIVITIES:

- establishment of an AI/ML specialized branch repository (instance);
- customization of the graphical user interface of the Clarin-DSpace system;
- testing and piloting of the repository;
- provision of access to researchers from the DM4AI group and improvement of the system based on their feedback (pilot phase);
- DM4AI will start using the repository for its own data; **AI_1**
- transfer of know-how, creation of tutorials. **AI_1.11**

2/ Ensuring of effective data retrieval/meta-data

We will create a metadata model that will be AI/ML[82] domain-specific and follow the NMD (integration with the National Metadata Directory is necessary for the provision of mandatory metadata to the NMD). The model will cover all aspects of interoperability that are necessary for further seamless use or sharing. The metadata model will still be maintained and regularly updated within the project to reflect all key properties of the stored data. The creation of a metadata schema will be accompanied by a series of analyses and studies that will cover: (i) information on existing schemas, (ii) identification of the scope of needs and (iii) determination of the structure of the schema. **AI_1.1**

We program SW to manage data and search capabilities and integrate it into the repository as an extension of the functionalities provided by Clarin-DSpace. The repository will thus facilitate the efficient retrieval of data/metadata, so that relevant data can be easily identified and prepared for large-scale learning or inference. Specialized data structures that reflect the type of query requests and the structure of data/metadata are a prerequisite for effective data retrieval capabilities. Examples of these include categorical data, time series, graphs, image and video data, or 3D point and network clouds. The repository will be one of the types of stored objects (for replicability of experiments) to support a description of the entire AI/ML workflow, including received data/medadata, applications for AI/ML training (e.g. source code, binaries, containers – versioning), trained models, and results. **AI_1.2**

ACTIVITIES:

- creation and maintenance of a domain-specific metadata model for AI/ML;
- ensuring of interoperability for sharing and using the metadata model;

[80] https://www.lexis.tech. LEXIS Platform – an advanced platform that provides easy and secure access to high-performance computing resources, including a web interface for distributed data management, running HPC applications, and orchestrating complex workflows across multiple sites.

[81] The training courses will be realized in cooperation with the EOSC CZ Training Centre, which will provide adequate capacities, technical facilities and organizational support and are not included in the budget for OS II.

[82] Artificial intelligence (AI) / machine learning (ML).

- data management and support for effective data and metadata retrieval; **AI_1.1**
- definition of criteria for the effective identification of relevant data;
- selection of suitable structures for different types of data (categorical data, time series, graphs, 3D point clouds...);
- optimization for effective search and querying;
- implementation of functions for storing and versioning AI/ML workflow. **AI_1.2**

3/ Support for activities necessary for the FAIRification of data in repositories, ensuring provenance and curation

We will use tools for generating and managing standardized provenance. These will be created as part of the PKA sub-activity Provenance, namely: 1/ SW for provenance storage and management; 2/ SW library for working with provenance according to CPM. The tools will be integrated with the repository and data transfer software according to the provenance of the methodology and will become part of them so as to enable the provision of machine-processable documentation of the origin of models and data in a standardized representation (link to W3C PROV standards[83], Common Provenance Model:ISO 23494[84], Workflow Run RO-Crate[85]).

No software solution currently exists that would allow the storage, management and use of AI models and data in accordance with the aforementioned provenance standards. Therefore, it cannot be assumed that the complete documentation of AI models (trained model -> dataset entering model training and testing -> "source" data set -> samples from which the sets were generated -> original sample source) will be interoperable. Facilitation of the existence and retrievability of interoperable and trustworthy provenance that would document the entire lifecycle of models and related datasets, with an emphasis on ensuring their confidentiality and privacy of related persons.

A methodology will be developed containing procedures for compliance with FAIR principles for research data, taking into account the specifics of the artificial intelligence domain (ethical issues, bias, complexity), i.e. methodology of data preparation, annotation and curation.

The methodology will describe the preparation, annotation and curation of the received data from the perspective of FAIR principles with regard to the needs of AI/ML workflow and related data. This will include, for example, methodological procedures on how to use tools that combine machine learning with a human verification loop or automatic duplication removal capability.

Currently, various tools, initiatives and methodologies have been created to serve as a guide for data creators, on the basis of which FAIR principles can be implemented. However, the vast majority of them work on a general rather than domain-specific basis. Tools such as FAIR Wizard[86] or FORCE 11[87] work only with general elements of FAIR principles and do not take into account the specific aspects of individual domains. However, specific FAIR principles for the social sciences (Data Management Expert Guide[88]) and natural sciences (RDMkit[89]) are available. A comprehensive methodology in terms of FAIR principles with regard to the needs of AI/ML workflow and related data is not yet available in the domain.

[83] https://www.w3.org/TR/prov-overview/.
[84] https://www.bbmri-eric.eu/news-events/common-provenance-model-bbmri-led-iso-ts-23494-12023-released/.
[85] https://www.researchobject.org/workflow-run-crate/.
[86] https://fair-wizard.com/.
[87] https://force11.org.
[88] https://dmeg.cessda.eu/Data-Management-Expert-Guide.
[89] https://rdmkit.elixir-europe.org/.

ACTIVITIES:

- definition of requirements related to the verification of the reusability of AI models in the context of provenance;
- analysis of the possibilities for addressing requests from the previous point in the repository using the PKA outputs of the Provenance sub-activity;
- design of architecture for the technical integration of PKA outputs of the Provenance sub-activity;
- implementation of the architecture from the previous point, pilot operation, demonstration of use; **AI_1.3**
- definition of key principles and standards for the preparation, annotation and curation of data in the field of AI/ML;
- development and formulation of methodological steps and recommendations;
- description of data preparation procedures in accordance with FAIR principles;
- documentation of the methodology and its distribution to target groups. **AI_1.10**

4/ Design of AI model licensing mechanisms

In the field of licensing, the activity will be phased as follows: (i) legal/technological analysis, (ii) development of a set of model licensing agreements, (iii) creation and pilot implementation of a metadata model, (iv) inclusion and verification of AI specifics over sensitive data, (v) validation and modification of proposed solutions.

We will analyse the nature of AI models from a legal and technological point of view. The resulting document will summarize the legal and technological aspects of AI models in the context of classification as software, databases, or a special type of intellectual property. It will contain an overview of relevant legislation (European and national), licensing mechanisms and recommendations for practice. **AI_1.4**

We will create Licence Models and model contracts for AI Models. A uniform set of licence models and model contracts (e.g. for multi-partner projects, working with students, dealing with sensitive data) allows for a clear definition of rights and obligations in the creation, sharing and use of AI models, reflecting at the same time the requirements of the GDPR[90] and trade secrets.

As part of the project, we focus on creating a concept of AI model licensing that takes into account the European legal context and current trends in the protection of computer programmes, datasets and model results. The project identifies the possibilities for licensing the source codes of AI models (e.g. GPL[91], Apache, mit[92], BSD[93]), protection of training parameters in the form of databases (under Directive 96/9/EC and licences such as ODbL[94] or PDDL[95]), as well as specific licences adapted to AI models, such as BigScience OPEN RAIL[96] or Model Openness Framework[97] (MOF) concepts.

The developed concept is oriented towards two main scenarios: a situation where multiple partners within or across institutions contribute to the development of AI models, and a situation where ready-made AI models are shared towards end users and other developers. Particular attention is paid

[90] General Data Protection Regulation (GDPR).
[91] General Public License (GPL).
[92] Massachusetts Institute of Technology.
[93] Berkeley Software Distribution (BSD).
[94] Open Data Commons Open Database License (ODbL).
[95] Planning Domain Definition Language (PDDL).
[96] https://bigscience.huggingface.co/blog/the-bigscience-rail-license.
[97] https://isitopen.ai.

to domains that work with sensitive data, where AI models can directly contain these data. **AI_1.5**

We integrate license models and model contracts with the metadata model (Metadata Model for Describing Licensing aspects of AI Models), where aspects related to licences must be included in the metadata. The Metadata Model for Describing Licensing Aspects of AI Models will be a machine-readable metadata model capturing licensing and legal information, including aspects of sensitive data or copyright protection. The model will be designed so as to be compatible with FAIR principles as well as with national and European standards (e.g. continuity with the National Metadata Directory). **AI_1.6**

ACTIVITIES:

- legal and technological analysis of the nature of AI models; **AI_1.4**
- model AI model contracts to other users;
- development of a set of model licensing agreements for AI models; **AI_1.5**
- creation and pilot implementation of a metadata model;
- integration and verification of AI specifics over sensitive data;
- validation and modification of proposed solutions;
- metadata model describing the licensing of AI models in a machine-readable manner, following the analysis and model licensing agreements;
- metadata model describing the data sources used for model training. **AI_1.6**

5/ Validation of created metadata models and deposition methodologies using AI models processing sensitive medical data

A further activity will be the validation of created metadata models and deposition methodologies developed initially for AI systems commonly trained on open data. Pilot deployment of these models on sensitive data – or a combination of sensitive and open sources – will allow verification of their usability, reliability and performance, even when processing protected medical or organic food data. The solution will include close cooperation with the TKA SENSI expert groups and the TKA B/H/F working team, ensuring that the validation mechanisms fully reflect the specific aspects of both domains. As part of the pilot verification, there will be no research, but only validation of architecture and metadata models connecting sensitive data and sensitive or potentially sensitive AI models; it will use existing resolved or live problems. We will utilize the existing pipeline of the RationAI research group at MU, which successfully develops models with deployment in clinical practice, along with sensitive data from multiple different medical institutions.

The activity will primarily validate metadata models and deposition methodologies, as well as provide input for their development, and thus does not require separate sustainability beyond the sustainability of models and methodologies. A secondary output will be an extended RationAI pipeline for processing clinical data for the training and inference of AI models; this will continue to be available as an open-source solution.

We will analyse open data sources and repositories at national and European level, focusing on both the technical possibilities of machine access to them and the conditions of access via an external service. An analysis and piloting of the connection of AI models to sensitive data sources and their processing in adequately secured environments (especially SensitiveCloud) will also be carried out.

ACTIVITIES:

- definition of the main objectives of the activity, requirements for the validation of metadata models and deposition methodologies, and identification of key stakeholders;

- input data analysis: identification and categorization of sensitive medical data to be used for validation, including image, time and tabular data;
- integration of the rationAI[98] pipeline: implementation of the RationAI pipeline for the processing and analysis of clinical data in conjunction with AI models;
- sensitiveCloud deployment[99]: use of SensitiveCloud infrastructure for the secure storage, transmission and processing of sensitive medical data;
- development of validation scenarios: definition of test cases and a methodology for the validation of metadata models and deposition methodologies using AI models;
- testing and validation: performance of test runs, debugging of AI models and iterative validation of metadata models;
- documentation and reporting: preparation of output reports and documentation in accordance with the requirements for documenting the output;
- publication of results: publication of results, including validated metadata models and deposition methodologies in an open repository (SW validation of created metadata models and deposition methodologies using AI models processing sensitive medical data). **AI_1.7**

6/ Analysis of possibilities for the integration of the repository with external data sources

It is not currently possible to easily use and search data from various European sources for the needs of the Czech AI/ML community. It is necessary to personally search the various repositories and then obtain data using the methods of the relevant repositories. We will perform a broad analysis of various European repositories, Data Spaces and other sources at the European level, with a focus on the possibilities for their integration with the DM4AI repository. The aim is to simplify research activities of the Czech AI/ML community that require external data sources.[100]

A second analysis, of the possibilities of integration with data sources at the national level, will be performed. The analysis will focus on the possibilities for integration with data sources that provide open data of state administration, local government and other institutions at the national level (e.g. CHMI) in order to make these data available for enriching the activities of the Czech AI/ML community with non-research data.

It is not currently possible to easily retrieve data from various open national data sources. It is necessary to search and browse these data sources individually and use the methods for downloading data that are particular to them. An analysis of selected open national data sources and the possibility of their integration with repositories from Activity 2.2 will be carried out.

ACTIVITIES:

- gathering and evaluation of existing data interfaces at European level;
- identification of key requirements and standards for integration;
- analysis of available technologies and tools for the integration of data interfaces;
- development of conclusions and recommendations for the integration of data interfaces at European level; **AI_1.8**
- identification of relevant data providers that provide open data to government, local government

[98] https://rationai.fi.muni.cz.
[99] https://www.cerit-sc.cz/infrastructure-services/sensitivecloud.
[100] Examples of initiatives with which the group envisages close cooperation: BDVA/DAIRO (https://bdva.eu/dairo/), CLAIRE (https://cairne.eu/), ELLIS (https://ellis.eu/), EUDAT (https://eudat.eu/), BBMRI (https://www.bbmri.cz/) a BigScience (https://bigscience.huggingface.co/).

and other institutions at the national level (e.g. CHMI[101]);
- identification of key requirements and standards for integration with these sources;
- analysis of existing tools and technologies for integrating data sources;
- development of conclusions and recommendations for the effective integration of data sources at the national level. **AI_1.9**

SUB-ACTIVITY OUTPUT CODES[102]

AI_1	DM4AI Repository
AI_1.1	AI/ML metadata model
AI_1.2	Data management and search options
AI_1.3	Integration of SW for generating and managing provenance with an AI repository
AI_1.4	Analysis of the nature of AI models from a legal and technological point of view
AI_1.5	Licence Models and model contracts for AI Models.
AI_1.6	Metadata model to describe licensing aspects of AI models
AI_1.7	SW – Validation of created metadata models and deposition methodologies using AI models processing sensitive medical data
AI_1.8	Analysis of the possibilities for integrating data interfaces at the European level
AI_1.9	Analysis of integration options with data sources at the national level
AI_1.10	Methodology of data preparation, annotation and curation
AI_1.11	Educational materials/tutorials

10.3.2. SUB-ACTIVITY 4.2 – INFERENCE ENVIRONMENT FOR GENERATIVE AI MODELS STORED IN THE DATA REPOSITORY

This activity will focus in particular on covering the specific needs of the Data Management for Artificial Intelligence and Machine Learning working group in the field of user interaction with generative AI models stored in the data repository. As the download and commissioning of these models requires considerable resources, it is desirable to allow limited interaction with them to allow greater familiarity with their potentialities and basic testing by users within the repository itself.

An environment for simple interaction with generative models will be created and linked to the data repository. Using specific metadata about a given model, it is installed on resources suitable for its launch in the correct software environment and made available to the user via a "chatbot"-type graphical user interface.

The Clarin-DSpace repository system will be enriched with the possibility of user interaction with generative data models for AI in the form of a chatbot with limited capabilities. This option will be a sort of equivalent of a "preview" of the dataset, because it will, like the preview of image data, provide a basic option to assess the suitability of a given model for users without having to download and run it on its own resources.

The following steps are planned for the activity.

First, we will analyse existing SW solutions for the effective inference of generative models, their

[101] https://www.chmi.cz/.

[102] The sub-activity output codes are linked to Annex 3 Schedule of Key Activities and chap. 11 Main planned outputs/products of Feasibility Study.

replaceability and practical tests of selected solutions and create a list of SW solutions of inference that the tool will use for maximum achievable compatibility with models and their formats and verified means of loading models. We will also analyse the transfer of data models from the repository to the inference environment, including the design of authentication and authorization for both the models in the repository and the inference environment.

We implement data transfers and automatic model launch on optimally selected hardware resources with maximum use of metadata in the dataset repository record. We further implement a graphical user interface for communication with models, including the implementation of a direct connection from the Clarin-DSpace repository.

The output of this will be SW that defines the environment for the inference of generative AI models stored in the data repository and interaction with them, with an emphasis on the efficiency of inference for different types and sizes of models. After requesting a "preview" of the model in the data repository, the model will be copied to the computing environment for inference, deployed on the appropriate hardware and software platform, and the user is informed that they can start communication with the model. A web interface will be available for this purpose.

The proposed environment will allow initial familiarization with the models without the user having to deal with these operational issues.

The main qualitative/quantitative benefits of the activity will be the shortening of the life cycle of the use of artificial intelligence and machine learning methods by optimizing data transfers and increasing user comfort by using a unified data management solution.

Thanks to the tool that has been developed, users will be able to test models on local infrastructure, automatically transfer them in the vicinity of computing resources and use large datasets. A common data management tool for the application of artificial intelligence and machine learning methods will also facilitate interdisciplinary cooperation. The quality of data will also increase with respect to FAIR principles thanks to the unified means of transferring data and its metadata. **AI_2**

ACTIVITIES:

- analysis of existing software solutions for the effective inference of generative models, their substitutability and practical tests of selected solutions;
- analysis of the transfer of data models from the repository to the inference environment;
- implementation of data transfers and automatic start of the model;
- implementation of a graphical user interface for communication with models;
- operation of at least one instance of the Clarin-DSpace repository system, which will have an integrated user environment for communication with generative models and will be functional and user-friendly throughout the entire chain, from requesting the launch of a compatible model from the repository to user communication with the model. **AI_2**

10.4. KA 5 – THEMATIC CLUSTER SOCIAL SCIENCES

PARTICIPATING PARTNERS: SI CAS (TKA Guarantor), CU, MU, UWB

We will create a domain repository platform for the social sciences based on the upgrade of two existing repositories: CSDA103 (Institute of Sociology) and DataHub104 (CU SCI) supplemented by a newly built repository for sensitive data. The effectiveness of achievement of goals is based on the connection of existing systems to the NDI EOSC in the Czech Republic and their supplementation with the necessary, previously missing element of the domain infrastructure. The comprehensive environment for the implementation of the Open Science policy in the social sciences aims at both the storage of data and their reuse in social science and inter-domain research. Existing systems are integrated into the international data services ecosystem; at the same time, the NDI will be connected to the European level of the domain data infrastructure.

The aim of upgrading the CSDA and DataHub repositories is to ensure full compatibility with the NDI and NMD environment, as well as the international infrastructure, integration of the platform into NDI, implementation of FAIR data principles, creation of links to the user communities of both data producers and data analysts, and the creation of new NDI services that will support data sharing using NDI and the use of innovative practices in research based on data sharing. There is no repository for sensitive data in social sciences in the Czech Republic. The construction of such a repository is necessary in this research area, in which working with sensitive data is unavoidable. Without such a repository it will not be possible to store and make available sensitive social science data in the Czech environment, and it will not be possible to ensure the implementation of FAIR principles. The efficient design of the new repository is based on the maximum use of the NRP environment and the results of TKA SENSI.

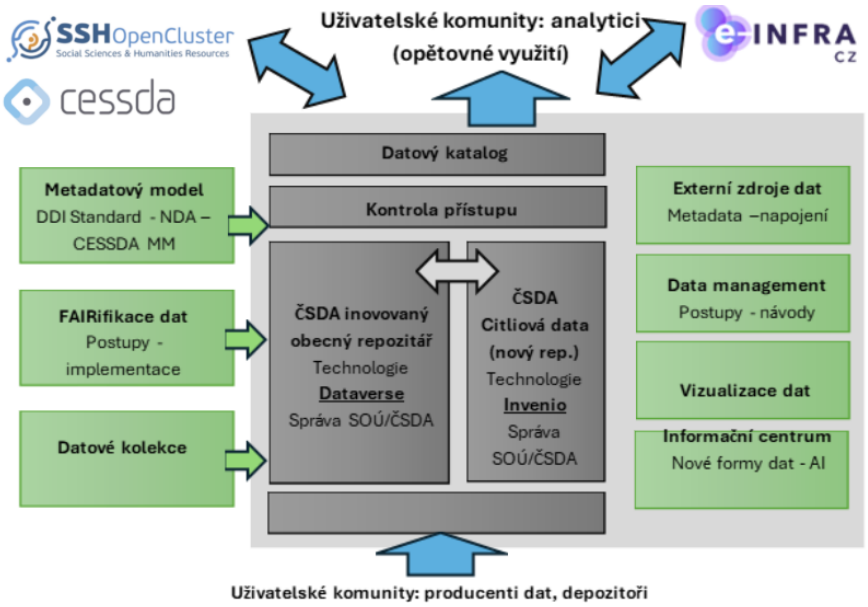

Fig. 6: Upgraded CSDA repository and newly created repository for sensitive data

Fig. 6 summarizes the interconnection of project activities to the repositories of the CSDA national data services centre. The upgraded CSDA general repository will be moved to the NRP environment. Data

103 Czech Social Science Data Archive (CSDA), Institute of Sociology of the CAS, https://archiv.soc.cas.cz/.
104 DataHub, Map and Data Centre, Charles University Faculty of Science, https://datahub.natur.cuni.cz/.

will be managed and cared for by the Institute of Sociology using Dataverse technology and will be functionally connected with the newly built CSDA – Sensitive Data repository. The FAIRified data of the existing repository will be integrated into the NDI. These activities will include the creation of baselines for connection to NDI and NMD (metadata model, FAIRification procedures) and the FAIR-ification of significant data collections. The storage and provision of access to primary data will be complemented by other NDI services. External data sources will be connected to the system (integration of metadata into the data library, annotation, data linking). Maximum access to these resources is also a priority for research communities in relation to addressing current societal challenges in the decision-making sphere. "Guidelines" on data management issues will directly support the use of platform services, data sharing standards and their quality, and the enforcement of Open Science policy. The implementation of data visualization tools will expand user communities, increase the information value of data (creation of time series), and contribute to the use of NDI. The Information Centre will support the promotion of innovative methodological procedures (new forms of data, AI).

The system will ensure the connection to the ESFRI105 research infrastructure CESSDA ERIC,106 and thus to the107 Social Sciences & Humanities Open Cluster (SSHOC) ecosystem of European data services and the European-level EOSC. Interconnection is based on the use of international standards and tools, in particular the international metadata standard DDI, CESSDA Metadata Model (CMM), Dataverse technology and the principles of the Open Archiving and Information System (OAIS),108 etc. The standard solution will also help in the process of obtaining the international Core Trust Seal (CTS) certification. The added value is to increase the quality of data and research and its international competitiveness, also with regard to the importance of international comparative research for scientific excellence in the field.

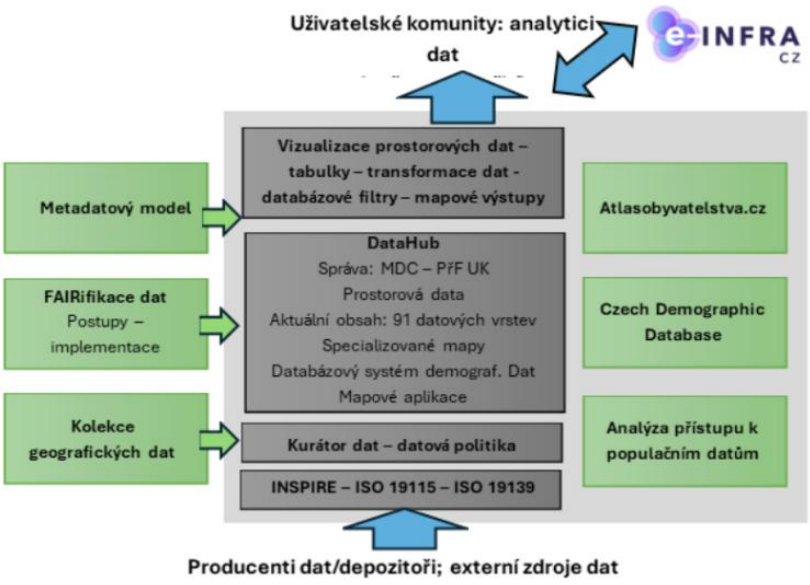

Fig. 7: Upgraded DataHub repository

105 European Strategy Forum on Research Infrastructures (ESFRI), https://www.esfri.eu/. Infrastructures included in the European Commission's Strategic Plan for European Research Infrastructures, known as the ESFRI Roadmap.
106 Consortium of European Social Science Data Archives European Research Infrastructure Consortium (CESSDA ERIC), https://www.cessda.eu/.
107 The project is also linked to other ESFRI infrastructures, namely ESS ERIC (European Social Survey, https://www.europeansocialsurvey.org/) and GGP (Generations & Gender Programme, https://www.ggp-i.org/).
108 *Open Archival Information System Model* (OAIS).

Fig. 7 describes the linking of activities to the upgraded DataHub repository, which provides data mainly in the field of social geography and demography in a specific environment for the presentation of spatial data. The system will undergo a significant update in order to connect to NDI and the NMD. Activities will include the creation of baselines, including a specific metadata model and FAIRification procedures, and the FAIRified collection of geographical data. We will also focus on the systematization and accessibility of demographic databases and population data through the system and newly created platforms (Atlasobyvatelstva.cz, Czech Demographic Database), thus responding to the long-term requirements of the research community.

The results and proposed partial solutions of the TKA SOC project will be shared with relevant domain partners for consultation, and will be presented at professional conferences both in the Czech Republic and abroad. In order to strengthen the presentation of the results, obtain feedback and consultation on the solution, a number of business trips abroad will be carried out, especially by members of the team of the Institute of Sociology of the Czech Academy of Sciences and the University of West Bohemia in Pilsen. It is expected that they will participate in professional conferences, workshops, seminars, or, if appropriate, summer or winter schools in the domains related to the project focus, i.e. the social sciences, work with professional data and information (e.g. data management, FAIR data), issues concerning participation in the EOSC, strengthening cooperation, etc.

We will also focus on the systematic dissemination of project outputs and the development of educational capacities in the field of data management, and not only for social science professionals. The goal is to raise awareness of new services, analyses and educational materials created in the project through a targeted communication strategy and a newly created web platform.

10.4.1. SUB-ACTIVITY 5.1 – UPGRADE OF THE CSDA REPOSITORY AND FAIRIFICATION OF DATA[109]

The Czech Social Science Data Archive (CSDA) currently makes data available mainly from sociological research, but also from research in some other fields (currently 742 data sets), including rich documentation (metadata) for secondary analysis in social science research. The project will expand the capacity of the data repository, and thus make it possible to store FAIRified data and make it available for other domains, such as research from experimental economics or new types of data, e.g. synthetic data, which enable innovative approaches to data analysis. The implementation of the activity will support both the expansion of communities using the repository and greater usability for existing communities, including support for the use of innovative, highly competitive data analysis methodologies. At the same time, a repository will be built for accessing sensitive data that cannot be stored in an insecure environment and made freely available.

The results of this project have a major impact on the field of experimental economics because it does not yet have a domain repository where data can be stored, nor developed methodologies for working with data. One specific aspect is the diversity of data generated within the field, be it by way of collection, laboratory, online, in the field or obtained from various software (e.g. Tree, oTree, Qualtrics) or physiological data (from eye-tracking, EEG, fMRI, etc.).

With regard to technical development, we will work on the use of AI tools and language models for working with data and possibilities for protecting the identity of respondents. This can be

[109] In the description of the KA, the planned activities are linked to the implemented outputs. The outputs are marked with a code. For TKA SOC, this is in the format **S_x** (key activity), **S_x.x** (partial, i.e. part of the key activity).

an important process for working with data, allowing it to be saved in a repository after research is completed and properly made available according to FAIR principles.

The newly emerging structure and capacities of the NRP and the existing CSDA system, built on the Dataverse platform for open data archives, will be used to upgrade the CSDA data repository. The system will be moved from the Institute of Sociology's server110 to the NRP environment. This will utilize advanced systems from the NRP environment (AAI, licensing, cybersecurity, replication system). The Dataverse system will be maintained, connecting the repository to (a) existing CSDA data services, (b) the CESSDA ERIC international data services environment, and through it to the European-level EOSC, including the SSHOC social science cluster, where CESSDA is the leading organization. All conditions and procedures will be set so as to ensure technical interoperability with NRPs and overall standardization within the NDI as a link to the NMD and NCR. In addition, the positions of repository administrator and data curator will be created to ensure the management, development and reception of data in the upgraded CSDA repository. **S_1**

We will prepare NMD-compatible metadata models in order to integrate the existing repository into NDI. A general metadata model for social sciences based on international domain standards will be developed for the purpose of accessing and processing various types of data from the social sciences. We are based on the international DDI standard 111, which is comprehensive, generic and continuously developed and commonly used in the social sciences. DDI will enable the acquisition of rich metadata of different types of data for the purposes of interoperability and secondary analysis in research. The project will define mandatory elements in terms of compatibility with the NMD metadata model and metadata models of the international data infrastructure (CESSDA CMM, etc.). **S_1.1**

Methodologies for working with sociological data and data from experiments in the field of economics will be developed. Both methodologies will deal with the management of research data, how to process them so that they can be stored in a repository, made available with an emphasis on FAIR principles, interoperability, and how to work with these data in secondary analysis. As data from sociological research, typically questionnaire surveys – quantitative, but also qualitative – are often used for secondary analysis, it is important to ensure that the data is described well and in detail. Experimental data from the field of economics represent a specific data type, the use of which required knowledge beyond that widespread in the field. The methodology will be used to expand this knowledge so that economic experimental data can be used in secondary analyses and replications. This is a pilot intended to verify the integration of a specific type of data into the data infrastructure in the field of social sciences. **S_1.3, S_1.4**

Several FAIR Implementation Profiles (FIPs) will be created for the FAIRification of data. These will define and specifically propose how to meet FAIR principles for a specific research community within the field of social sciences. **S_1.3, S_1.4**

We will share all upcoming methodologies and metadata models at the international level for consultation with regard to interoperability and ensuring the interconnection of both data and services in the international environment. For this reason, the results of the project will benefit from participation in international forums, conferences, seminars, dealing with data management, FAIR, interoperability, etc. It will also be important to establish close cooperation with major international research infrastructures and other important initiatives (e.g. EOSC, Research Data Alliance, CESSDA). **S_1.3, S_1.4, S_1.1, S_1.2**

110 Institute of Sociology of the CAS.
111 https://ddialliance.org/.

FAIRified datasets will be stored in the repository in collections of individual domains or major research, etc. The assumption is that a large number of new datasets will be stored in the CSDA repository as part of the project. For this reason, only significant collections are listed here (for data from experimental economics, sociology and synthetic data). Data already published in the CSDA repository will be modified and FAIRified. The necessary work on dictionary translations, data cleaning, metadata editing will be carried out for reasons of interoperability and FAIRification. The data will be provided with metadata according to the appropriate metadata model. For this purpose, translations of documentation (part of metadata), implementation of the ELSST international thesaurus, etc. will be implemented for the purpose of integration into international data services = FAIRification of data. The data will be checked and processed for the purposes of data deposition, provision and sharing. For data that is already stored, the metadata will be cleaned, supplemented or modified according to the required standard. **S_1.7, S_1.6, S_1.8**

A methodology for creating and working with synthetic data will be developed separately. Synthetic data in sociology arise as artificially created data that contain for researchers the same information about the society as the empirical data obtained, but no longer contain the original information from the original respondents. We can say that the original data from the respondents are mixed in such a way that the original respondents cannot be traced, but that it makes an equally valid statement about society as a whole. Based on real empirical data, new "virtual" respondents, who never existed but behave statistically the same as those in the original dataset, are simulated. It is verified that synthetic data correspond to the characteristics of real empirical data, but at the same time do not disclose information about specific individuals. They are also used in testing analytical methods, algorithms or tools without harming any research or respondents. They are also often used in teaching and training. They allow you to simulate hypothetical scenarios, for example, you can create a model of a company with certain characteristics and see what would happen if the unemployment rate or the electoral system changed. As this is a new type of data, there are no procedures for working with these data from the point of view of data management, their FAIRification, or accessibility in data repositories. These data are difficult to create (both financially and in terms of time), so sharing it is beneficial for scientific and research work. The information necessary to develop the methodology will also be obtained through consultations with experts, expert forums and feedback when presenting partial results. **S_1.5**

With the development of artificial intelligence, the risk of revealing the identity of respondents increases even in cases where data is formally anonymized. Modern AI tools can deduce or guess the identity of specific people from a combination of seemingly anonymous data (e.g. age, gender, location, occupation). This issue is highly topical and will be fundamental in the future, but is currently not being addressed sufficiently. The analysis will focus on the use of AI tools for the application of processes that can anonymize or deanonymize social science data. AI resources will also be used to analyse the possibilities for revealing the identities of respondents using AI tools and to describe the possibilities for the protection of data against identity disclosure, with an emphasis on the possibilities for their reusability. This will include a critical search of freely available resources that can be misused to identify respondents. We are therefore pursuing two objectives: 1) to verify, with the aid of AI, whether it is safe to continue to share data that we consider to be anonymized and de-identified by current standards, or may lead to the identification of respondents when combined with other publicly available data; 2) to find, with the assistance of AI, ways to reliably anonymize and de-identify data. **S_1.10**

Most sets of quantitative and qualitative data are scientifically useful and reusable even if they do not contain personal and/or sensitive data. Therefore, it is necessary to analyse the appropriate means of anonymizing datasets containing personal and sensitive data, which will allow the publication

of data files that originally contained personal and sensitive data. Using language model services, a procedure will be proposed to anonymize sensitive qualitative data so that they can be used for secondary analyses. An additional source for obtaining information to create the analysis will be consultations with experts and presentations of partial results in professional forums, which will provide feedback. **S_1.9, S_1.10**

ACTIVITIES:

- connection of the CSDA repository to the NRP and its development and operation; **S_1**
- preparation of a general model for social sciences and a metadata model for demographic data; **S_1.1 and S_1.2**
- creation of several FAIR Implementation Profiles (FIPs); **S_1.3, S_1.4**
- development of a methodology for the possibilities of creating synthetic data and working with synthetic data; **S_1.5**
- FAIRification of data collections (for experimental economics, sociological data, synthetic data); **S_1.6, S_1.7, S_1.8**
- analysis of the possibilities for the process of encoding sensitive text and its subsequent publication without sensitive data; **S_1.9**
- analysis of the possibilities for revealing the identities of respondents using AI tools and describing the possibilities for the protection of data against identity disclosure, with an emphasis on the possibilities for their reusability. **S_1.10**

SUB-ACTIVITY OUTPUT CODES[112]:

S_1	connection of the CSDA repository to the NRP and its development and operation
S_1.1	General metadata model for social sciences
S_1.2	Metadata model for demographic data
S_1.3	Methodology/standard for working with data from sociological research
S_1.4	Methodology/standard for working with data from experiments in the field of economics
S_1.5	Methodology of possibilities for creating synthetic data and working with synthetic data
S_1.6	Collection of FAIRified data for experimental economics
S_1.7	Collection of FAIRified data: sociological data
S_1.8	Collection of FAIRified data: synthetic data
S_1.9	Analysis of the possibilities for the process of encoding sensitive text and its subsequent publication without sensitive data
S_1.10	Analysis of the possibilities for revealing the identities of respondents using AI tools and describing the possibilities for the protection of data against identity disclosure, with an emphasis on the possibilities for their reusability

10.4.2. SUB-ACTIVITY 5.2 – CREATION OF A NEW CSDA REPOSITORY FOR SENSITIVE DATA AND FAIRIFICATION

In the Czech Republic there is currently no infrastructure for storing and sharing sensitive data in the social sciences in a way that meets the requirements of the evolving Open Science environment,

[112] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

legislation and ethical standards. The activity will create this infrastructure, verify its functionality and integrate it into the domain data services environment in the EOSC-CZ and European CESSDA, SSHOC and EOSC ecosystems. The research community will have the ability to store and share these data in a standard way that is consistent with FAIR data principles and the requirements of Open Science, in a broader multidisciplinary and international environment. The activity will bring a qualitative shift for the effective use of data in qualitative and other research based on the analysis of sensitive data. At the same time, it will raise the standards of protection of research subjects (GDPR and ethics) and ensure high security standards, including cybersecurity. This will also have an impact on the possibilities for international cooperation (data availability, comparison), and thus the competitiveness of research. This will create an environment suitable for accessing sensitive data of the public sphere, which is a priority requirement of both the community and research policies. The project includes the storage of and access to sensitive data (previously 0, now min. 10 datasets).

A new repository will be created for sensitive data, i.e. data that require special protection for the personal data of research subjects and other data requiring special protection. The repository will be created using the INVENIO software system with appropriate modifications for sensitive data. e. The Institute of Sociology/CSDA will ensure curatorial administration and determine the conditions of access and settings in cooperation with the NRP. Furthermore, the role of the repository administrator, which will be closely involved in the preparation of the repository and then in its management, will be established. The principles of data archiving, data policy and CSDA archiving rules will be developed or upgraded on the basis of implementation. Access Control settings will be made established on the outputs from TKA Sensitive Data and according to domain standards. A modified general metadata model, described above, will be used to describe the stored data. **S_2, S_1.1**

The social sciences often generate knowledge about society based on the processing of individual personal data. It is not always possible to process and analyse these data in anonymized form. The archiving and disclosure of such data either requires de-identification of the data subjects while maintaining the required informative value of the data (methods of data anonymization and minimization of the risk of identity disclosure). If this is not possible, specific procedures and organizational and technical measures should be put in place to secure personal data when they are used in research. A methodology will be developed on how to prepare sensitive data for the repository and how to disclose them. During development, the methodology will be continuously shared for consultation with professionals and will incorporate partial outputs of TKA SENSI. **S_2.1**

The proposed methodology will be verified when processing model datasets with sensitive data in a newly built repository with a high degree of access control and specific procedures for deciding on its allocation. In order to test data deposition, processing and distribution (accessibility) processes, data that are already available to the data archive will first be used; however, these data have only very limited options for disclosure. Existing data will be modified and FAIRified, and in particular a comprehensive check of the data content will be carried out, including an analysis of the risk of identity disclosure (disclosure risk assessment). A FAIR Implementation Profile (FIP) that defines how to apply and comply with FAIR principles in the deposition, processing and distribution of this type of data in the social sciences will be prepared. Next, the metadata will be checked and prepared according to the corresponding metadata model. The new repository will be registered in the NCR. We anticipate that ten datasets with sensitive data will be published at the end of the project. **S_2.2**

ACTIVITIES:

- building a repository for sensitive data and its operation; **S_2**
- creation of a methodology for data management, the archiving and sharing of sensitive data in social sciences, and access to sensitive data in the repository; **S_2.1**
- testing of data deposition, processing and distribution processes. **S_2.2**

SUB-ACTIVITY OUTPUT CODES[113]:

S_2	New sensitive data repository for the social sciences
S_2.1	Methodology for data management, the archiving and sharing of sensitive data in the social sciences, and the disclosure of sensitive data in the repository
S_2.2	Pilot collection of sensitive data and metadata to be stored in the repository

10.4.3. SUB-ACTIVITY 5.3 – UPGRADING OF DATAHUB REPOSITORY AND FAIRIFICATION OF DATA

The DataHub of the Map and Data Centre serves as the central platform of the Charles University Faculty of Science for the sharing, visualization and analysis of geographical, and especially spatial, data. It offers tools for working with data, open datasets, interactive map applications, analytical dashboards and a map viewer. It gives access to scientific information from the fields of social and physical geography and demography. Spatial data are provided in the form of layers, geodatabases, and formats of the geographic information system (GIS). DataHub currently contains 91 data layers/geodatabases. It utilizes the standard ArcGIS Online environment with the ArcGIS Hub extension (enabling the creation of a geoportal with a metadata catalogue), that is used by the commercial, but international, geographic community. The licence is paid for by Charles University and utilized by many workplaces. It is expected that it will be used in the long term. The creation of one's own portal environment outside of ArcGIS would require an order of magnitude greater investment in technical facilities and qualified staff. All activities leading to the connection of the DataHub repository to the NDI/NMD are feasible under the existing licence in combination with open procedures, and do not require additional costs. As this is a specific type of data, it cannot be made available in a common data repository (or only to a limited extent and without added value spatial information that can be displayed in the integrated map application). Nevertheless, these are important data sources, so their FAIRification and access to metadata in the NDI environment will be beneficial not only for the research and scientific community, but, due to its form, for the general public too.

At the same time, the possibilities for accessing population data, which are an important part of a number of research projects, and integration into NDI, will be explored.

The DataHub platform will be extended to include connection to the NDI/NMD system. We will address the automation of data recording, obtaining of identifiers (persistent link, DOI), and metadata management using a custom programme that will use the REST API. This will be accessible to the administrator and data curators of the Map and Data Centre, who will oversee the recording and management of data. Connection to the NDI common authentication and authorization infrastructure will be implemented. System logs, standards and data policies will be upgraded based on implementation. Furthermore, a database system will be programmed to allow the expansion of services to include specific

[113] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

data outputs from the field of demography (non-spatial database data). The positions of repository administrator and data curator will be created to ensure the management, development and reception of data in the upgraded repository. **S_3**

Metadata models for geographic and demographic data that ensure compatibility with this system will be prepared for the integration of the existing repository into the NDI/NMD. Furthermore, the repository will be registered in the NCR. In order to ensure interoperability, development of the first model will take into account the specific aspects of geographical data and the requirements of the GIS environment. It will be based on international standards regarding geographical data – ISO 19115, 19139 and INSPIRE. A search will be carried out for international standards and norms that can be used as the basis for handling demographic data. **S_1.2, S_3.1**

Geographic data that are or will be published in the upgraded DataHub repository will be FAIRified.

For reasons of interoperability and FAIRification, it will be necessary to work with the translation of dictionaries, data cleaning, and metadata editing. Data that is already stored (91 data layers) will be cleaned and metadata will be added in accordance with the required standard. The data will be checked and processed for the purposes of data deposition, provision and sharing. It is anticipated that more than one dataset will be made available in the course of the project. **S_3.2**

The availability of detailed population data is necessary for many areas of research (e.g. demography, sociology, epidemiology, economics or regional planning), as it allows the construction of indicators and the performance of analyses that cannot be carried out on the basis of commonly published aggregated data. Compared to other countries, there is much room for improvement in the Czech Republic with regard to access to microdata (i.e. data on individuals in the populace, such as age, gender, education, etc. at the individual level, and not aggregation within groups) from administrative sources and in the possibilities for their use. The analysis of international standards and approaches to the provision of detailed population data and microdata will be applied to the Czech Republic, including recommendations aimed at simplifying access to the necessary data, their integration into the NDI environment and expansion of the possibilities of their use in accordance with the principles of Open Science. The information necessary to develop the analysis will also be obtained through consultations with experts, expert forums and feedback when presenting partial results. **S_3.3**

ACTIVITIES:

- connection of the DataHub repository to the NRP and its development and operation; **S_3**
- ensuring of compatibility with the repository system (metadata models); **S_3.1, S_1.2**
- FAIRification of the collection of geographical data of the Czech Republic; **S_3.2**
- analysis of international standards and methods of providing population data. **S_3.3**

SUB-ACTIVITY OUTPUT CODES[114]:

S_3	Upgraded DataHub repository
S_3.1	Metadata model for geographic data
S_3.2	Collection of FAIRified data: detailed geographic data of the Czech Republic
S_3.3	Analysis of international standards and methods of providing population data.

[114] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

10.4.4. SUB-ACTIVITY 5.4 – DEVELOPMENT AND PILOT IMPLEMENTATION OF SERVICES AND TOOLS FOR THE SOCIAL SCIENCES

This activity is the culmination of the construction of a domain repository platform within the EOSC ecosystem in the Czech Republic based on the construction and integration of extension services of the data infrastructure. Based on the results of the above activities, specific CSDA and DataHub services for the research community will be built and connected to the national and international EOSC ecosystem. Tools for data visualization (time series analysis) will be created and implemented. Completely new services for DataHub in the field of demographic data (Atlasobyvatelstva.cz) will be created. Emphasis will also be placed on important domain sources, such as data from international large research infrastructures with Czech participation and the possibilities for sharing and utilizing them. All services will expand the possibilities for interdisciplinary use and increase or expand their usability.

A visualization software web tool of social science data from sample surveys showing the development of attitudes, opinions and beliefs of the target population over time will be created. The tool will facilitate the filtering of results by socio-demographic and other substantive indicators. This will make it possible to monitor the dynamics of the development of individual groups within the population. The tool will focus on one of the basic methods of the social sciences. Time comparisons will expand the user group of data services beyond analysts who know statistical methods (i.e. they will also bring data services closer to a number of experts in the decision-making sphere). The tool will be purposefully developed to be flexible, with the possibility of supplying additional measured concepts and sorting characters. In the pilot operation, it will be deployed on time series from the CSDA repository from the data of the Centre for Public Opinion Research (due to the high use of CVVM data, the reach to the target group will be increased), with the possibility of extending to other social science data sources. **S_4**

The integration of the Atlasobyvatelstva.cz platform for specialized maps into DataHub data services will be carried out. These are unique data that are usable in other fields. As part of the previous activity, they will be FAIRified, but at the same time a special interface is required to display and work with them; this will be developed as part of the project. The Atlasobyvatelstva.cz platform for specialized maps will thus be revitalized. **S_5**

In the project, we will expand the existing DataHub (CU SCI) with an innovative new database focused on detailed demographic data of the Czech Republic. The database will offer harmonized, high-quality data on fertility, mortality, migration and population structure at different geographical levels (Czech Republic, regions, districts) and timelines. Its free availability and support for advanced analyses will make it a key source for demographic and interdisciplinary research, including the creation of documents for state administration and self-government. Integration will require extra programming work to effectively access and display the data. **S_6**

Due to the fact that secondary data analysis is a common activity in the social sciences, knowledge of data sources is important. In accordance with the Project Call, landscape analyses focusing on significant data producers and data sources will be performed in order to increase the use of repositories. Data sources whose data cannot or should not be placed directly in the repository platform will be selected. These will be made available to users as follows: a) harvesting of available metadata, FAIRification and transfer to the metadata model of the repository platform, inclusion in the CSDA catalogue (Dataverse) and publication, linking to the data source (data can be obtained from the original source), b) creation of descriptions, overview and Web guide to data sources completely outside the repository

platform, which will be published on the CSDA website. This will expand the services of the repository for the needs of the research community**. S_7**

An analysis focusing on ways of sharing and the practical use of data from international research infrastructures such as ISSP, EVS, SHARE ERIC, ESS ERIC and GGP, with an emphasis on how data can be obtained for research and study purposes, will be carried out. As these international infrastructures also carry out data collection within the Czech Republic, they are a very important resource for research. Therefore, documents will be developed to provide clear instructions and supporting materials on how to gain access to international data, how to work with them and how to use them in research. These will be published on the CSDA website and will become an integral part of educational activities. The analysis will also include an assessment of access policies, licensing regimes and examples of good practice that can serve as inspiration for data sharing. Ways will be sought to establish closer cooperation within the constraints that currently exist in terms of disclosure rules. The issues and possible solutions will be discussed with experts in professional forums. **S_8**

ACTIVITIES:

- creation of a visualization software web tool for social science data; **S_4**
- integration of external data sources; **S_7**
- integration of the Atlastobyvatelstva.cz platform for specialized maps into DataHub; **S_5**
- extension of the existing DataHub with a demographic database integrated into data services; **S_6**
- analysis of sharing and practical use of data from international research infrastructures. **S_8**

SUB-ACTIVITY OUTPUT CODES[115]:

S_4	Data visualization tool
S_5	Atlasobyvatelstva.cz platform for specialized maps
S_6	Demographic database integrated into DataHub data services
S_7	Integration of external data sources
S_8	Analysis of sharing and practical use of data from international research infrastructures.

10.4.5. SUB-ACTIVITY 5.5 – EDUCATIONAL PLATFORM AND COMMUNICATION STRATEGY, DISSEMINATION OF PROJECT OUTPUTS WITHIN THE PROFESSIONAL COMMUNITY IN SOCIAL SCIENCES

The subsequent activities focus on the systematic dissemination of project outputs and the development of educational capacities in the field of data management, and not only for social science professionals. The goal is to raise awareness of new services, analyses and educational materials created in the project through a targeted communication strategy and a newly created web platform. The activity supports practical skills and theoretical knowledge in the field of data work in accordance with FAIR principles and responds to the need for systematic support in working with various types of social science data. It significantly contributes to the development of data culture, supporting interdisciplinary cooperation.

The project will achieve a number of unique outputs that are usable for professionals in the social sciences. This activity will aim at expanding the outputs (newly established services, educational

[115] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

materials, elaborated analyses) to the Czech social science community. The aim of the proposed activity is the implementation of a targeted communication strategy and dissemination of project outputs within the professional community in the fields of social sciences (stakeholders, depositors, members of user communities, researchers, project managers, etc.). This is related to the dissemination of information about newly established services, created analyses and produced educational materials and signposts and pointing to their practical usability. The implementation of this strategy is based on the team's long-term experience and contacts in the professional community. Its implementation will be connected to the provision of wide-ranging services to social sciences professionals. The targeted communication strategy will result in a significant increase in awareness of data services and will also contribute to improving the culture of data sharing and management in the Czech social sciences through a series of dissemination activities that will emphasize wider outreach within user communities, as well as interdisciplinarity in the social sciences (e.g. increasing awareness of FAIR data and data reusability and notification about user-centred educational activities in the main areas of data management).
In order to implement the strategy, a special section of the CSDA website will be set up, through which some communication with stakeholders will take place (texts with information about new services, publications of materials created in the project, promotions regarding project activities, etc.). The content and form of the strategy will also target stakeholders in the social sciences (appropriate selection of language and professional terminology, use of existing communication channels). The proposed communication strategy will be implemented in connection with other project activities, will reflect the interests and needs of the main target groups, and its contribution will be subject to continuous evaluation during the project. **S_9**

Furthermore, a domain portal will be created to support education in the field of DMP (Data Management Plan) and data management. This educational web platform will include educational materials related to the management of various types of social science data. On this platform, a general tutorial on data management of social science data and several sets of educational materials for individual types of data (sociological data and sensitive data, synthetic data, demographic data, population data, data from experimental economics), which will be based on the outputs of other activities of this project, will be created. Furthermore, they will be built on the professional research of existing materials for social science data management and on the knowledge and experience of project team members who teach social science data management at HEI and specialized workshops. The platform will also include a signpost of relevant professional resources for data management in the social sciences.

In the materials published through the platform, emphasis will be placed on the main areas of data management, i.e. data retrieval for secondary analysis, data planning in a research project (DMP), data documentation and processing, ethical and legal aspects of data production and data work, data protection, short-term data storage and long-term data archiving, and curating and public data sharing. Knowledge of data management is not primarily technical (control of software for data management), but practical and theoretical. It is essential that data producers (researchers, project managers) and users of secondary data have knowledge of data management, which will enable them, in cooperation with data repositories and other stakeholders, to produce and use data that meets FAIR principles.

The web platform containing the educational materials described above will be primarily intended for
a) researchers who produce and/or scientifically use data for their own research, and b) students in

the social sciences. It will also be useful for teachers of data management and the methodology of social science research as a basis for teaching. **S_10.1, S_10.2**

Following the creation of analyses, methodologies and other materials on the issue of new forms of data and the use of AI in research as part of other activities of this part of the project, sets of information materials, resources and procedures that will be included in the structure of ČSDA services, or even DataHub, will be made available on the website in a systematic and comprehensive form.

This information centre in the form of an educational web platform will support the development of new, state-of-the-art practices in social science research, open up important issues in the field of ethics, technology and the legal environment, and provide information on proven solutions. At the same time, it will be a starting point for the broader involvement in international cooperation of research infrastructures in this area. No systematically created infrastructure currently exists in this area**. S_10**

We plan to organize several training courses in connection with the educational materials created. These will be mainly intended for data providers, students, researchers, teachers and other users of social science data services.[116]

SUB-ACTIVITY OUTPUT CODES[117]

S_9	Communication strategy and dissemination of project outputs within the professional community in the social sciences
S_10	Information centre on issues affecting new forms of data and the use of AI in research into data services
S_10.1	Materials for data management education and methodology of social science research
S_10.2	Creation of guidelines for data management education and methodology of social science research

10.5. KA 6 – THEMATIC CLUSTER PHYSICAL SCIENCE

PARTICIPATING PARTNERS: JHIPC CAS (TKA Guarantor), CU, UPOL

1Contemporary physical science relies on ever larger and more diverse volumes of data – from massive data streams of particle physics at the LHC (already over 1.5 petabytes of data per day) and in DUNE or Pierre Auger projects, series measurements in material and microscopy laboratories reaching up to terabytes in volume. The EOSC European initiative, the Czech National Data Infrastructure (NDI) and the forthcoming legislation on open access require that these data are **FAIR** – i.e. Findable, Accessible, Interoperable and Reusable. Without building tools and standards, there would be a risk of losing the scientific and economic potential of these unique ensembles and the Czech Republic would not be able to fully participate in international infrastructures and projects with a long-term perspective.

KA 6 is therefore **building a specialized branch repository known as "Physics" based on the implementation of the Invenio system** in the National Repository Platform (NRP). The repository will offer robust a storage area, clearly defined metadata models and automated tools for the mass transfer

[116] The training courses will be realized in cooperation with the EOSC CZ Training Centre, which will provide adequate capacities, technical facilities and organizational support and are not included in the budget for OS II.
[117] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

and verification of large data packages from the ATLAS ITk, DUNE, CTAO SST-1M, Auger and crystallographic structural analysis experiments. This ensures that data generated today will be stored in accordance with FAIR principles from the very beginning and will be traceable and usable over the long term for future generations of researchers.

The technical section is connected to **innovative services**:

- an **automated metadata creation** tool linked to Electronic Laboratory Notebooks (ELNs);
- a web component for the **direct visualization of multidimensional data** (HDF5/NeXus, fits, etc.) in the browser;
- and **e-learning modules and materials for workshops** that will rapidly expand FAIR data management skills among both Czech and international teams. These open-source tools will also be immediately usable in other NDI clusters and will strengthen the interoperability of the entire Czech infrastructure.

Key experiments are entering the phase of upgrade or live operation at this time (ATLAS Run 4 in 2030, first data from DUNE in 2031). If we set up the storage architecture, access rights management and FAIR processes during the preparation, we will save considerable financial resources (in the order of millions of CZK) for later data conversion, reduce the risk of data loss and gain a competitive advantage when participating in ESFRI and Horizon Europe projects. In addition, by creating a unified physical repository, we will ensure that Czech researchers have a strong negotiating position when sharing data and computing resources with international partners, and support the transfer of knowledge to industry, medicine and energy.

With this cluster, we meet the strategic goal of the Call: to cover the entire life cycle of physical data – from its creation through standardized storage and opening to the world to targeted community education. The result of this will be a sustainable ecosystem that turns voluminous physical data into a truly reusable source of innovation for science and society.

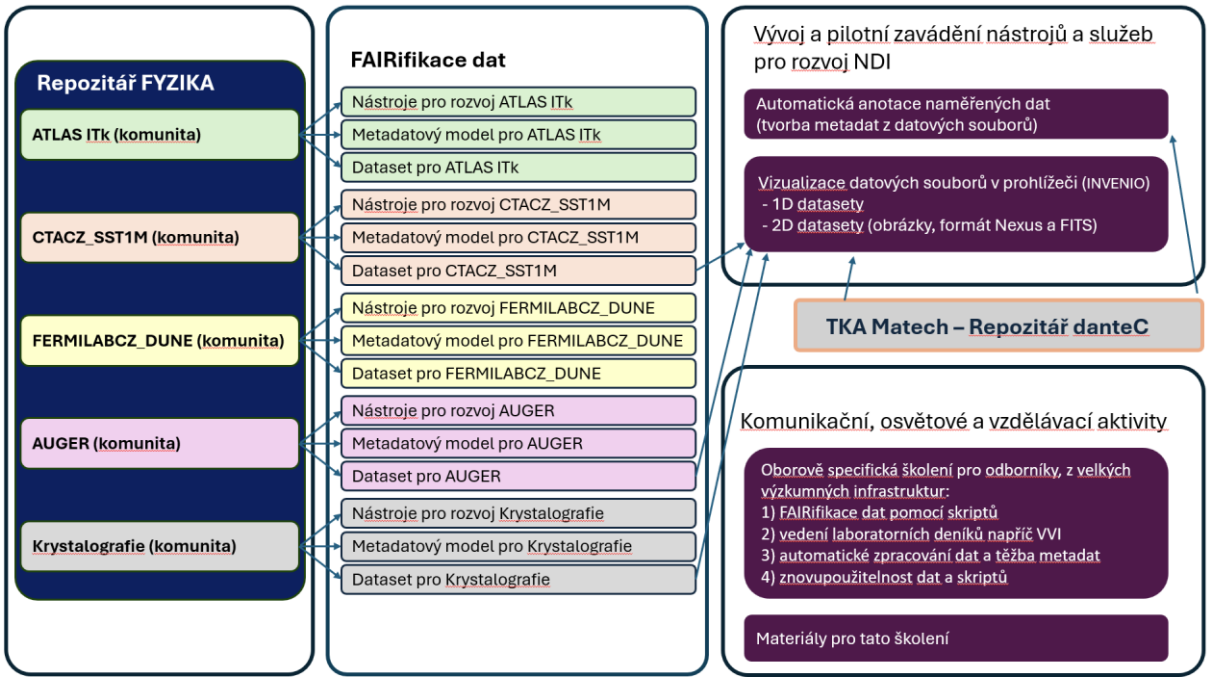

Fig. 8: Graphical display of TKA PHYSICS

10.5.1. SUB-ACTIVITY 6.1 – CREATION OF A NEW REPOSITORY OF PHYSICS AND DATA FAIRIFICATION[118]

Following the NRP project, we will create a new repository for the domain cluster Physics. The repository will utilize the repository service provided and managed by the NRP project. It will be a standard implementation provided by NRP on the Invenio platform. The sub-activity will ensure the functionality, security and usability of the repository for users who will enter data, as well as for users who will access data. Specific documentation for this repository will be prepared with specific examples of data input procedures and their description using metadata. **F_1**

The internal structure of the Physics repository will be organized with respect to the communities and datasets of new large physics experiments:

- **ATLAS ITk**: Newly built inner tracker (ITk) of the ATLAS experiment (A Toroidal LHC ApparatuS) on the Large Hadron Collider of the European Organization for Nuclear Research (CERN);
- **FERMILABCZ_DUNE**: DUNE (Deep Underground Neutrino Experiment) experiment at FermiLab in the USA;
- **CTACZ_SST1M**: Functional telescope prototype for the Cherenkov Telescope Array Observatory (CTAO);
- **AUGER**: Pierre Auger Observatory (PAO);
- **Crystallography**: Advanced crystallographic structural analysis (ASTRA, i.e. Advanced Structure Analysis).

This corresponds to the individual logical sub-units, which are described in detail below.

With an emphasis on FAIRification, appropriate metadata models, tools for automatic data input, tools for verification of stored data and methodologies for the use of data in further research will be created as part of the activities of the named sub-units.

We expect that TKA PHYSICS will thus create models for the development of Open Science in the wider physical community, especially in connection with the projects of large research infrastructures of the Road Map of the Czech Republic, through which the Ministry of Education, Youth and Sports supports open access to unique facilities.

At international level, this will facilitate participation in the Roadmap of the European Strategy Forum on Research Infrastructures (ESFRI). Specific examples of the intended dissemination are the COMPASS Upgrade program and the PALS Research Centre (Prague Asterix Laser System), on which the Institute of Physics cooperates with the Institute of Plasma Physics of the Academy of Sciences of the Czech Republic, as well as Extreme Light Infrastructure (ELI).

The MGML (Materials Growth & Measurement Laboratory) and CzechNanoLab research infrastructures, which cooperate closely with TKA MATECH, are also directly involved in TKA PHYSICS through their parent institutions the Institute of Physics and CU MATHPHYS.

The success of TKA PHYSICS also depends on the adoption of established communities and developed tools at the international level, i.e. in large experiment collaborations. Therefore, the results will be presented and offered to users in the form of active speeches, which are planned, for example, at the following conferences:

[118] In the description of the KA, the planned activities are linked to the implemented outputs. The outputs are marked with a code. For TKA PHYSICS, this is in the format **F_x** (key activity), **F_x.x** (partial, i.e. part of the key activity).

- Accelerating the Adoption of Open Science (CERN, Geneva);
- Open Science Fair: Fusing Forces – Accelerating Open Science through Collaboration (CERN, Geneva).

While domain-focused, the reach of these conferences is considerable; this is also due to the size of collaborations.

In addition to the aforementioned domain-focused conferences, we will also provide outreach to the wider community in the physical sciences, for example at the TRIPLE (Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration), "Improving Discovery and Collaboration in Open Science" or similar conferences.

1/ ATLAS ITk

We will create a new community in the Physics repository with specific metadata for sensor data for ATLAS ITk. We will transfer already recorded data stored on a local disk array to the Physics repository. The data will be converted to a suitable format and described with metadata. Access to data will be reserved for the necessary period for members of the ATLAS collaboration only, whose priority is justified by their participation in the construction or operation of the ATLAS experiment. The approach solution will utilize the AAI tools developed by the NRP project. The virtual ATLAS organization that registers authorized workers already exists is used for grid computing. In cooperation with NRP, we will develop and implement solutions to utilize the existence of this Virtual Access Control Organization in the national repository. **F_1.1**

A laboratory at the Institute of Physics is analysing sensors for the upgrade of the internal ATLAS detector, preparations for which are currently underway for its commissioning with the start of Run 4, planned for 2030. The data obtained, with a total size of several tens of terabytes, must be stored for a long period of time (for the duration of data acquisition using these sensors and the analysis of the data, i.e. until after 2040) to allow the retrospective inspection of the quality of individual sensors. The data includes numerical measurement results and photographs taken by the sensors. The collection of FAIRified data will be entered into the Physics repository for the needs of the domain community. Specific metadata will be generated for these data. To this end we will create a metadata model and a suitable web interface or scripts for accessing the data, as well as automated tools for uploading and verifying them. **F_1.2, F_1.1**

We will utilize tools for the mass transfer of data from a local disk server to the repository. As they will have to deal with a large number of files with a total size of several tens of terabytes, these tools will have to be robust enough to produce consistent records in the repository even in the event of sudden network outages, problems when reading from local storage areas, or the instability of some repository components due to congestion or other reasons (Invenia version updates, server maintenance, operating system upgrade, etc.). If possible, the tools will utilize existing software libraries developed by the NRP project or other groups so that maintenance of the software does not require a large amount of effort when changing the API of the repository, or other changes.

2/ FERMILABCZ_DUNE

The DUNE (Deep Underground Neutrino Experiment) will measure the properties of neutrinos produced in the accelerator in the Fermilab laboratory. A group comprising the Institute of Physics and the Institute of Computer Science of the Czech Academy of Sciences, and the Charles University Faculty of Mathematics and Physics, is participating in the project through the Fermilab-CZ research infrastructure. The giant detector will consist of four modules, each containing 17,000 tons of liquid

argon. 300,000 silicon photomultipliers (SiPMs) will be used to read the signal. These photon detectors are currently being measured in several laboratories so that they can be installed for the planned start of data collection in 2031. The results of measurement are currently stored on disk servers at the measuring apparatus. The properties of such a large sample of SiPMs are of considerable interest, and can also be used for other projects. The results of the measurement will be described by metadata, stored in the new repository community and made available on the basis of the open data policy of the DUNE project. As it is planned for data acquisition to last at least until 2042, the data must also be stored reliably for the study of the long-term reliability of the SiPMs.

We will create a new community in the Physics repository with specific metadata for data from silicon photomultiplier measurements for the DUNE experiment. It transfers already recorded data stored on the local repository in the laboratory setup to the repository. A dataset will be created from the measurement data of silicon photomultipliers. The data will be recorded continuously until the end of the project in 12/2028. The data will be converted to a suitable format and described by metadata in accordance with FAIR principles. Access to the data will be reserved for members of the DUNE collaboration team and for the necessary period of time only. The approach solution will utilize the AAI tools developed by the NRP project. A DUNE Virtual Organization (with staff from more than 35 countries) that registers authorized staff already exists and is used for grid computing. In co-operation
with the NRP, we will develop and implement solutions on how to utilize the existence of the DUNE Virtual Organization for managing access in the national repository. **F_1.4, F_1.3**

Using the tools/services provided by the NRP project, we check the consistency of the datasets and assign them the required access rights. After processing the existing data, we will continuously add new records from new measurements.

We will need tools to bulk convert the data to a suitable format and size and to transfer them from the local storage area to the repository. As they will have to deal with a large number of files, these tools will have to be robust enough to produce consistent records in the repository even in the event of sudden network outages, problems when reading from local storage areas, or the instability of some repository components due to congestion or other reasons (Invenia version updates, server maintenance, operating system upgrade, etc.). Where possible, we will use pre-existing software developed by the NRP project or other groups so that the maintenance of these tools does not require a large amount of effort when changing the API of the repository or other changes.

3/ CTACZ_SST1M

ERIC CTAO (Cherenkov Telescope Array Observatory) is building an extensive network of telescopes in two locations (on the Canary Islands in Spain and in the Atacama Desert in Chile) to observe gamma rays coming from space. Together with other partners, the Czech group has prepared two prototypes of the SST-1M (SST – Small-Sized Telescopes, mirror area diameter 4 m) and operates both on the grounds of the Astronomical Institute in Ondřejov. The volume of data collected has already exceeded 100 TB. We will develop a metadata model for these data, create a new community in the general repository of Physics, and store the data. With the help of AAI experts, we will establish access rights for sharing, first for members of the collaboration team only, with the plan to fully open up the data after the embargo expires. We will hold consultations on and share the created metadata model with the entire CTA collaboration (which is preparing other types of telescopes) and explore the possibilities of storing selected data for the entire collaboration. **F_1.5, F_1.6**

We will create a new community in the Physics repository with specific metadata for SST1M telescope measurement data. This will transfer to the repository and FAIRify data that have already been

recorded and stored on the local repository and on backups in the object repository provided by the CESNET Data Repositories service. The data will be converted to a suitable format and described with metadata. Access to the data will be reserved for members of the SST-1M collaboration team and for the necessary period of time only. The approach solution will utilize the AAI tools developed by the NRP project. **F_1.6**

4/ AUGER

The Pierre Auger Observatory has been collecting data from gamma-ray–detecting silicon photomultipliers since 2004, with collection planned to continue past 2030. It gradually releases these data for public access in an appropriate format. The observatory does not yet have a permanent repository for primary measurement data or for a larger collection of data from Monte Carlo simulations. Our group will ensure the FAIRification and storage of these valuable data, the collection of which involves a large international collaboration of 500 employees from 18 countries, in a domain repository. This entails the creation of a new community with an appropriate description using metadata based on a specific metadata model for data and simulations (in accordance with the PAO specification). The measured data will be gradually transferred and FAIRified from the existing iRODS-based storage area in the CC IN2P3 computing centre in Lyon. Due to the number of files, it is necessary to use tools for reliable data transfer and data integrity check in the target repository. The measured data will be supplemented with simulations using the Monte Carlo method. Since the volume of simulated files is very large (>1 PB) and they are stored in a distributed way in various centres involved in the European Grid Infrastructure (EGI), it will be necessary to select data suitable for long-term storage and those data to which the published results refer. Pierre Auger's collaboration gradually releases the measured data to the public. In cooperation with AAI specialists, access to various data sets will be managed based on the open data policy of the Auger project (licensing terms are governed by CC BY-SA 4.0 standards). **F_1.8, F_1.7**

We will use tools for mass transfer of data from the grid storage area to the repository. The data is currently stored in the iRODS (Integrated Rule-Oriented Data System) system in CC IN2P3 in Lyon, but it is planned that it will be moved to the grid centre in Bologna. Therefore, the tools will have to support different input storage areas.

5/ Crystallography

The ASTRA (Advanced Structure Analysis) department is part of CzechNanoLab and provides structural analysis of crystalline and polycrystalline materials for users from all over the world (the number of structures determined is close to 1,000 per year) through Open Access. The structures of both inorganic and organic samples, including biologically significant structures and drugs, are determined using both X-ray diffraction and electron beam diffraction methods. The department is also the originator and administrator of JAN's analytical software, which is a global standard for the analysis and refinement of structures.

Structural analysis is an example of a field that traditionally maintains a database of acquired structures, has a standardized CIF (Crystallographic Information File) data format, and whose data is validated before publication. A large number of structures exist in multiple recognized databases, including the Open Access Crystallography Open Database (COD). However, the original data are usually not available due to their size (ASTRA stores the original data on 100 TB disk arrays). Measured sample

data are managed through the ASTRA online system[119], which, however, does not meet current FAIR standards and does not provide access to data through persistent data identifiers.

Crystallography is an area of physical research that differs in approach from previous sub-activities. Unlike large collaborations involving hundreds to thousands of researchers working together on large-scale experiments, crystallography is used by a large number of smaller research groups. In addition, it typically represents one of the many approaches used to study material samples, and without microscopic characterization, spectroscopic and dedicated experiments, it would only make limited sense. Therefore, it also represents that component of physical research that is related to material research, the main challenge of which is to link data from various experiments and material simulations through the linked data approach; cf. TKA MATECH.

We will prepare suitable conditions for the interdisciplinary crystallography community and the metadata model in the Physics domain repository. The main activities will focus on FAIRification (to the maximum extent allowed by the existing metadata structure) of datasets of original measurements of structural analysis. We will develop appropriate tools for mass data transfer and prepare a metadata model for their description. **F_1.10, F_1.9**

We will develop tools for mass data transfer and tools for the selection and visualization of crystallographic data.

SUB-ACTIVITY OUTPUT CODES[120]

F_1	Creation and operation of a repository for the Physics domain
F_1.1	Metadata model for ATLAS_ITk
F_1.2	New ATLAS_ITk community
F_1.3	Metadata model for FERMILABCZ_DUNE
F_1.4	New FERMILABCZ_DUNE community
F_1.5	Metadata model for CTACZ_SST1M
F_1.6	New CTACZ_SST1M community
F_1.7	Metadata model for AUGER
F_1.8	New community for the AUGER observatory
F_1.9	Metadata model for Crystallography
F_1.10	New community for crystallographic data of the Institute of Physics – Crystallography

[119] https://astra.fzu.cz/sampleman/public/login.

[120] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

10.5.2. SUB-ACTIVITY 6.2 – DEVELOPMENT AND PILOT IMPLEMENTATION OF SERVICES/TOOLS FOR NDI DEVELOPMENT

The repository tools of the Physics thematic cluster respond to the needs of research teams working with experimental data in the physical sciences, and especially in laboratories, where datasets of complex structures and various formats are created. The starting point for this is a situation in which scientific data is stored electronically, but often without the necessary context, systematic metadata or links to experimental workflow. This reduces their reusability and long-term interpretability.

Through two sub-activities, tools will be developed for the systematic resolution of this problem: 1) automatic generation of metadata during measurements and their interconnection with the Electronic Laboratory Notebook (ELN); 2) direct visualization of data stored in repositories in a common web browser environment.

The resulting components will support interoperability within the National Data Infrastructure (NDI), will be adapted to physical data formats such as HDF5, NeXus or FITS, and will be developed as open-source in order to serve other research communities.

Automated metadata generation tool F_2

The aim of this partial sub-activity is the development and pilot launch of a tool for automated generation of metadata from measurements of physical quantities and their interconnection with the Electronic Laboratory Notebook (ELN). It is a domain-specific tool for experimental material research in physics. It reduces the administrative burden on researchers, increases the quality of metadata, and promotes interoperability within NDI.

The tool will provide:

– automatic login of researchers to the ELN (connection to AAI);
– retrieval of sample information via QR codes;
– extraction of metadata directly from the output data sets of measuring instruments (e.g. HDF5, NeXus, CSV, ASCII).

The system will be connected to the open-source Python Research Infrastructure User System (PRIUS) and developed as a modular system to allow its adaptation to various types of experiments. The tool will be implemented in the large MGML research infrastructure at CU MATHPHYS.

ACTIVITIES:

- design of system architecture and linking to ELN, AAI and repository APIs; **F_2.1**
- development of metadata parsers and their integration with ELN; **F_2.2**
- connection of the system to the QR identification of samples; **F_2.3**
- creation of user and installation documentation; **F_2.4**
- pilot use of the tool at workplaces of CU MATHPHYS
- finalization and publication of the tool in an open-source repository.

We will design the overall architecture of the tool for the automatic generation of metadata from measurements. The architecture will include components for integration with the Electronic Laboratory Notebook (ELN), connection to the AAI infrastructure (e-INFRA CZ) for user authentication,

and an interface enabling the export of metadata to repositories within the NDI. The proposal will be discussed with the developers of the PRIUS system and the developers of the Invenio platform within the NRP. **F_2.1**

We will develop a set of parsers that facilitate the extraction of metadata from commonly used data formats in experiments (e.g. CSV, HDF5, NeXus). These modules will be integrated into the ELN environment so that they automatically pre-fill relevant fields from data files. Emphasis will be placed on modularity, scalability and the possibility of adaptation to various experimental protocols. **F_2.2**

We will launch a system for identifying samples using QR codes. Every sample will be marked with a unique QR label for the automatic assignment of samples to measurements. This function will be designed with an emphasis on user-friendliness and continuity with sample registration within institutional databases. **F_2.3**

We will prepare documentation covering the installation, configuration and normal use of the system. The documentation will include technical requirements, deployment scenarios in different environments, and instructions for administrators and end users. It will also include demonstrations of data inputs and outputs with regard to FAIR standards. **F_2.4**

We will carry out the deployment and operational testing of the system in the real environment of two laboratory workplaces of the MGML infrastructure at CU MATHPHYS and the Institute of Physics of the Czech Academy of Sciences. All key functions will be verified, including the accuracy of the generated metadata, linking to AAI, and linking to sample records. We will collect feedback from users in order to fine-tune the tool before its release.

After the pilot verification, we will finalize the tool and release it as an open-source package. It will be ready for easy deployment in other infrastructures that use ELN or similar systems. The code will be published with an open licence.

Tool for the visual display of data files F_3

The aim of the partial sub-activity is the development of a tool for the visual display of datasets of measurements of physical quantities directly in an Internet browser. The output will be an effective and user-friendly way for NDI users to view research data without the need to install specialized software. The partial sub-activity meets the needs of domains that work with multidimensional data and supports data reusability in accordance with FAIR principles. In addition, the tool will make data viewing accessible to the general public, which does not have specialized tools for viewing it.

We will develop a tool based on modern web technologies (Vue) that will enable the display of data files in HDF5/NeXus (1D and 2D), FITS (2D image data) and Metadata4Ing format sessions. The tool will support interactive zoom, selection of data slices and the display of basic metadata, and will be integrated into the INVENIO repository APIs. The code will be distributed as an open-source component suitable for integration into NRP repositories and domain instances.

We will design the structure and appearance of the web interface, with an emphasis on clarity for users in the physical sciences. The interface will allow the display of different types of data and will be optimized for common browsers without the need to install additional software. The architecture will separate the presentation layer from data processing, making it easier to maintain and extend functions. **F_3.1**

We will develop components that facilitate the display of physical data stored in HDF5 and NeXus formats. Users will be able to browse the file structure, select 1D data series (e.g. spectra) or 2D image

data (e.g. maps, diffraction patterns) and edit the display (scales, colour schemes, cut-outs). Visualization will be built on libraries such as h5web or Plotly. **F_3.2**

We will ensure the display of 2D image data in the FITS format, which is often used in astrophysics, and other imaging methods. The tool will facilitate basic interactive work with images. Emphasis will be placed on ease of use, even for users outside the astronomical community. **F_3.3**

We will develop a module for the visualization of relations between individual experimental steps according to the Metadata4Ing metadata format. We will utilize visualization libraries (e.g. D3.js) to display experimental workflows in the form of interactive graphs, enabling clear orientation in the data and their context. **F_3.4**

The tool will be deployed in the Physics **F_1** and DANTE[c] **M_1**repository environments.

Upon completion of development and testing, the tool will be published as open-source software with open documentation. Source code, installation scripts and sample data will be available. The tool will be ready for use by other repositories on the Invenio platform.

ACTIVITIES:

- design of user interface and technical architecture; **F_3.1**
- implementation of support for the HDF5 and NeXus formats; **F_3.2**
- implementation of support for FITS; **F_3.3**
- implementation of support for Metadata4Ing; **F_3.4**
- deployment and operation at CU MATHPHYS;
- publication of the tool as an open-source package with documentation.

SUB-ACTIVITY OUTPUT CODES[121]

F_2	Tool for the automated generation of metadata
F_2.1	Design of system architecture and links to ELN, AAI and repository APIs
F_2.2	Development of metadata parsers and their integration with ELN
F_2.3	Connection of the system to the QR identification of samples
F_2.4	Creation of user and installation documentation
F_3	Tool for the visual display of data files
F_3.1	User interface and technical architecture design
F_3.2	Implementation of support for HDF5 and NeXus formats
F_3.3	Implementation of support for fits
F_3.4	Implementation of support for Metadata4Ing

10.5.3. SUB-ACTIVITY 6.3 – CREATION OF E-LEARNING COURSES AND MATERIALS FOR EDUCATION

The sub-activity responds to the specific needs of researchers and technicians in physics, especially those working at large research infrastructures. In the physical sciences, it is common for infrastructures (e.g. synchrotrons, neutron sources, large detectors) to automatically publish raw data to repositories, often without sufficient context or methodological support for their FAIRification and further

[121] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

use. This situation places increased demands on users who have to supplement the data with metadata, structure it and ensure its comprehensibility and long-term interpretability.

E-learning courses will therefore offer an overview of procedures for automated processing of data from large research infrastructures, including the creation of data management plans (DMPs) that will be linked to real data flows and Electronic Laboratory Notebooks (ELNs). Course content will also include work with scripts for the FAIRification and reuse of data over the long term, as is typical for physical experiments. Training events focused on the advanced use of ELNs across infrastructures and direct work with EOSC tools will be implemented in parallel to this. Outputs will be prepared in connection with the training framework of the EOSC National Training Centre and included in the NDI educational offer[122].

We will create a series of interactive e-learning modules that will be available online and cover the following topics:

- processing and management of data from large research infrastructures;
- creation and maintenance of data management plans (DMP) using ELNs in large research infrastructures;
- use of EOSC tools for the management, FAIRification and sharing of research data.

Courses will be prepared with an emphasis on clarity for the target group, complemented by practical exercises, examples from research practice and instructions. The output will be fully functional and publicly available educational content that can be used even after the end of the project. **F_4.1**

We will create materials for four thematic workshops/training courses for researchers and technical staff at large research infrastructures. The topics of individual workshops will be:

- FAIRification of data using scripts;
- keeping of electronic laboratory notebooks in a multi-infrastructure environment;
- automatic data processing and metadata mining;
- long-term reusability of data and scripts.

Educational materials will combine theoretical interpretation with practical tasks and will be designed in such a way that participants acquire specific skills that are applicable in their institution. Materials created will be made available online.[123] **F_4.2**

SUB-ACTIVITY OUTPUT CODES[124]

F_4.1	Series of three e-learning courses for PHYSICS
F_4.2	Educational materials for four PHYSICS workshops

[122] The training courses will be realized by the EOSC CZ Training Centre, which will provide adequate capacities, technical facilities and organizational support.

[123] The training materials are an output of the OS II project. The training courses/workshops themselves will be realized in cooperation with the EOSC CZ Training Centre, which will provide adequate capacities, technical facilities and organizational support and are not included in the budget for OS II.

[124] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

10.6. KA 7 THEMATIC CLUSTER HUMANITIES AND ARTS

PARTICIPATING PARTNERS: ICL CAS (TKA Guarantor), MU, CU, UWB

TKA HUMA will focus on the upgrading of existing and the development of new repositories in the humanities cluster, as well as the development of tools and services for the development of the NDI. Depending on specific technical conditions and domain and institutional policy, when describing TKA, we define a repository as (a) a specific instance of a given repository system with the possibility of partial specifically defined units or (b) a partial logically defined unit within such an instance in the form of a specific community or collection with specific curatorial procedures, including any individualized variants of metadata models. The aforementioned repositories are concentrated around four existing large research infrastructures for the humanities as central producers of research data and competence centres in the field of open science within the cluster (the Archaeological Information System of the Czech Republic (AIS CR), the Czech Literary Bibliography (CLB), the Czech National Corpus (CNC) and LINDAT/CLARIAH-CZ). They will utilize existing repository solutions, which will be supplemented by new repositories or repository communities in selected fields with significant research data production.

The implementation of the OS II project will consist of a total of five sub-activities. The individual sub-activities will be linked to a total of four repositories, each of which will focus on one of them: upgrades will be implemented in the (1) LINDAT/CLARIAH-CZ repositories (DSpace system; operated by CU MATHPHYS), which will expand their user community with the field of corpus linguistics by connecting the data of the CNK research infrastructure, on the basis of which a separate collection/community of the repository will be created, and the LINDAT/CLARIAH-CZ repository will be supplemented with FAIRified datasets of cooperating institutions (CU MATHPHYS, CU ARTS). The CU ARTS repositories will also be upgraded, namely (2) the institutional Digitalia Muni ARTS repository and (3) the ArchaeoVault repository for archaeology (both on the Islandora system). The user communities of both these repositories will be extended to users outside the parent institution thanks to their links to the NRP. In addition, it is planned that the ArchaeoVault repository will coordinate activities with the existing AIS research infrastructure. A new (4) Repository for Bibliographical Data, operated by the research infrastructure of the Czech Literary Bibliography (Institute of Czech Literature), will be created. Depending on the results of internal analysis, this will take the form of a separate collection/community within one of the existing repositories, or a separate instance of one of the NRP repository systems (ASEP, DSpace, Invenio). The last sub-activity is (5) the development and pilot introduction of superstructure tools for TKA HUMA repositories, the content of which will be the preparation of a total of four such outputs to expand services and increase the user comfort of individual repository platforms existing at TKA HUMA (Czech National Corpus, CU ARTS, UWB).

The results of TKA HUMA will be shared for consultation with relevant domain partners and presented at professional conferences both in the Czech Republic and abroad. Therefore, in order to strengthen the presentation of the project outputs, a number of business trips abroad are planned. These will be realized by team members from the Institute of Czech Literature of the Czech Academy of Sciences, CU MATHPHYS and CU ARTS and will involve either the presentation to partners abroad of the solutions used and consultations with them during exchange stays, and/or appearances at conferences at relevant professional events. At the same time, we expect the presentation of the project's outputs at prestigious international conferences, especially in the fields of (mathematical) linguistics or the digital humanities, as well as other disciplines in the humanities. Considering the TKA as a whole, we anticipate at least ten active presentations in the form of a conference paper or poster.

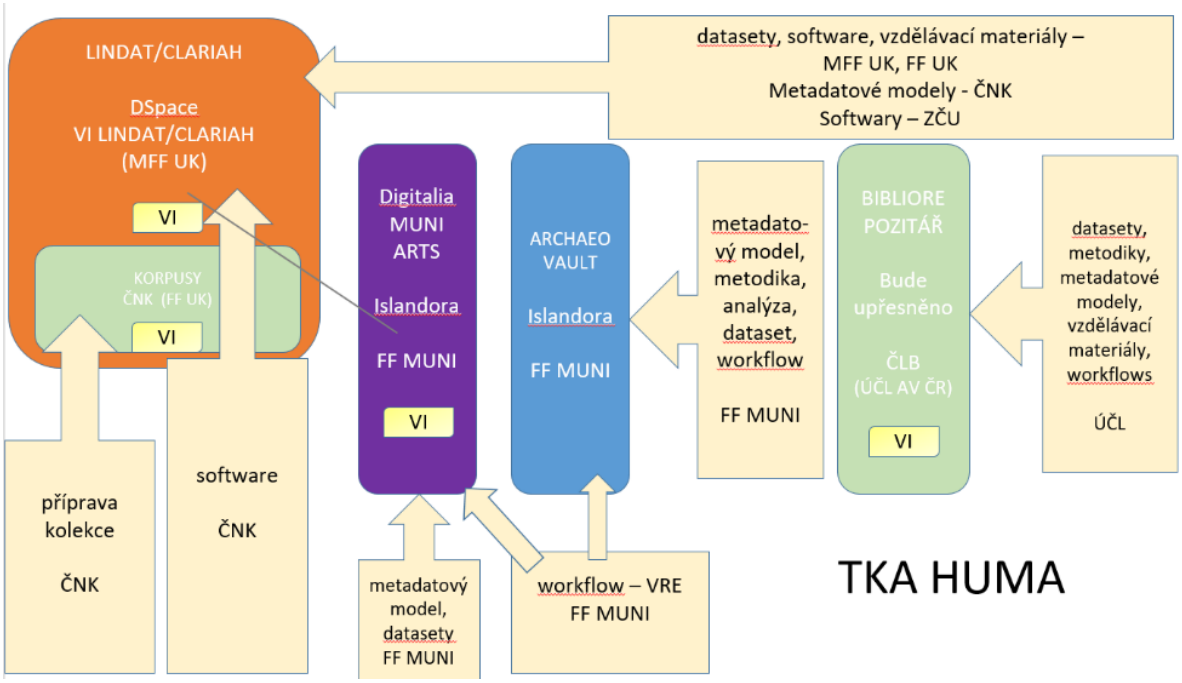

Fig. 3: Graphical diagram of activities

10.6.1. SUB-ACTIVITY 7.1 – UPGRADING OF THE LINDAT/CLARIAH-CZ REPOSITORY AND FAIRIFICATION OF RELATED DATA[125]

The LINDAT/CLARIAH-CZ repository has been operated by the research infrastructure of the same name in the DSpace system for a long period of time. The administration and development of the repository is continuously overseen by the Institute of Formal and Applied Linguistics at CU MATHPHYS. It is currently the largest and most comprehensive existing repository in the HUMA cluster. Following the expansion of the LINDAT/CLARIAH-CZ large research infrastructure, the original linguistic focus of the repository is changing and now includes other scientific disciplines and institutions that have become part of the LINDAT/CLARIAH-CZ consortium (a total of fifteen partner institutions), for which it offers the possibility of storing research data, or creating sub-domain or institutional collections/communities. As part of the OS II project, the LINDAT/CLARIAH-CZ repository will be upgraded by creating a specific collection/community of the Czech National Corpus (CNC), which will be curated by the research infrastructure of the same name, and the LINDAT/CLARIAH-CZ repository user community will thus also expand to include corpus linguistics as a cluster-important producer of domain data. For the purposes of integrating CNC data (and more generally for data of a similar type accessible via remote services), a separate metadata scheme will be developed that takes into account their specific characteristics so as to enable their easy integration into the repository. The content of the LINDAT/CLARIAH-CZ repository will be extended to include related outputs from the field of corpus linguistics, namely the FAIRified InterCorp dataset (CNC community). Other datasets of individual institutions of the LINDAT/CLARIAH consortium involved in the OS II project will also undergo FAIRification, namely selected datasets of CU MATHPHYS and CU ARTS. In addition to FAIRified datasets, a set of educational materials for working with the repository and a set of software for processing the exposed data will also be prepared to facilitate the involvement of both new and existing user communities. Thus, the implementation of the OS II project will enrich the LINDAT/CLARIAH-CZ

[125] In the description of the KA, the planned activities are linked to the implemented outputs. The outputs are marked with a code. For TKA HUMA, this is in the format **H_x** (key activity), **H_x.x** (partial, i.e. part of the key activity).

repository with an important user community of corpus linguists and at the same time they will undergo a fundamental evaluation, or, taking into account FAIR principles, significant domain datasets (including the InterCorp corpus, updated versions of the CU MATHPHYS linguistic datasets or the digital edition of the Wycliffe Bible) and software tools for their processing will be included in them for the first time.

The results of this activity will be presented at relevant domain conferences such as the Association for Computational Linguistic Conference (ACL), the International Conference on Language Resources and Evaluation (LREC), the Conference on Empirical Methods in Natural Languages Processing (EMNLP), etc. There are also plans for visits to leading domain workplaces engaged existing cooperation, during which the results of the project would be presented and further consultations held (the preliminary plan includes, among others, Oxford University, the University of Colorado, Brandeis University and Brown University). CU MATHPHYS is planning eight trips abroad as part of this activity. Methodological activities related to the operation of the LINDAT/CLARIAH-CZ repository and the implementation of Open Science in the humanities and arts will also be carried out as part of this subactivity.

Upgrade of the repository

In accordance with the Call, in the course of the project the repository will be (a) integrated with the NMD, to which mandatory metadata will also be provided, (b) connected to the EOSC common authentication and authorization infrastructure, (c) connected to the technical data transfer tools, methodologies and standards created by the EOSC-CZ IPs and the project ensuring the construction of NRPs, (d) registered in the NQF. Depending on the possibilities and domain practices, the repository will also be integrated into the international research environment. **H_2**

Development of a metadata model

A metadata schema for data accessible via remote services will be developed as part of the project. This is particularly necessary with regard to the specific situation of language corpora, which often cannot make source data available in the form of full texts or outside their own servers due to copyright. **H_58**

Preparation of materials for education

As part of the project, a set of domain-specific educational materials describing the practical possibilities of exhibiting and documenting individual datasets in the DSpace repository will be prepared in order to meet FAIR principles. **H_1.9**

Software development

As part of the project, a software package with converters and validators for individual PDT-C-EOSC (The Prague Dependency Treebank – Consolidated) datasets will be developed to ensure the consistency and coherence of the exposed data. **H_1.2**

FAIRification of datasets

As part of the project, selected key datasets of institutions involved in the research infrastructures of LINDAT/CLARIAH and the Czech National Corpus will be FAIRified. The data of individual datasets will be cleaned, harmonized, enriched with persistent identifiers if necessary, checked for consistency and integrity, and subsequently made available for use within the repository.

We anticipate FAIRification of the following datasets: Digital Critical Edition of the Wycliffe Bible, Lexical-Semantic Database of Czech, Database Manuscripta.cz – Database of Medieval Manuscripts

in the Czech Republic, PDT-C-EOSC (The Prague Dependency Treebank – Consolidated), Database CoCzeFLA – Corpora of Children's Czech in the Natural Environment, and Korpus InterCorp. **H_1.3–1.8**

ACTIVITIES:

- upgrade of the LINDAT/CLARIAH-CZ repository with the new Czech National Corpus collection;
- development of a metadata model, ensuring its availability in a public repository;
- preparation of materials for education, made available in a public repository;
- development of a SW package with converters and validators for individual PDT-C-EOSC FAIRification datasets available in the LINDAT/CLARIAH repository;
- methodological and support activities related to the operation of the LINDAT/CLARIAH-CZ repository and the implementation of Open Science in the environment of the humanities and arts.

OVERVIEW OF SUB-ACTIVITY OUTPUT CODES[126]

H_1	LINDAT/CLARIAH-CZ repository – involvement of CNC infrastructure language corpora as a separate repository collection
H_1.1	Development of a metadata schema for data accessible via remote services
H_1.2	Software package (LINDAT-KVS) with converters and validators
H_1.3	FAIRified dataset: Digital Critical Edition of Wycliffe's Bible
H_1.4	FAIRified dataset: Lexical-Semantic Database of Czech
H_1.5	FAIRified dataset: Database Manuscripta.cz - database of mediaeval manuscripts in the Czech Republic
H_1.6	FAIRified dataset: PDT-C-EOSC
H_1.7	FAIRified dataset: CoCzeFLA database - corpora of children's Czech in a natural environment
H_1.8	FAIRified dataset: InterCorp corpus
H_1.9	FAIR data during deployment and operation of the DSpace repository instance within EOSC CZ – educational materials

10.6.2. SUB-ACTIVITY 7.2 – UPGRADE OF DIGITALIA MUNI ARTS REPOSITORY AND FAIRIFICATION OF RELATED DATA

The Digitalia MUNI ARTS repository functions as an institutional repository of the Faculty of Arts of Masaryk University. The OS II project will include a fundamental upgrade, encompassing its connection to the NRP. This will significantly expand the existing user community, which will now include researchers outside the Faculty of Arts of MU. Procedures and tools for long-term archiving of existing data will be developed for these purposes. The content of the repository will then be enriched with three FAIRified MU ARTS datasets with fundamental value for research within the HUMA cluster. These will be the Dictionary of Czech Philosophers, a dataset of rare Diatheca photo-documents and a base of contemporary private correspondence. Interdisciplinary interoperability of the exposed data will be ensured thanks to the use of the metadata model for CIDOC-CRM cultural-historical data, which will be applied to selected pilot datasets.

[126] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

In accordance with the Call, in the course of the project the repository will be (a) integrated with the NMD, to which mandatory metadata will also be provided, (b) connected to the EOSC common authentication and authorization infrastructure, (c) connected to the technical data transfer tools, methodologies and standards created by the EOSC-CZ IPs and the project ensuring the construction of NRPs, (d) registered in the NQF. Depending on the possibilities and domain practices, the repository will also be integrated into the international research environment. The sub-activity will include methodological activities related to the operation of the Digitalia MUNI ARTS repository; Open Science will also be implemented in the humanities and arts. **H_2**

As part of the project, the CIDOC-CRM ontology will be applied to selected datasets of the Digitalia MUNI ARTS repository. These datasets will subsequently be made available in the RDF format, which will significantly strengthen their interoperability. **H_2.1**

As part of the activity, three other datasets from the provenance of the Faculty of Arts of Masaryk University will be FAIRified. With these datasets, the data structure will be revised, then they will be cleaned and checked for data consistency. The resulting files will then be made available in the repository. The relevant datasets are the Dataset Dictionary of Czech Philosophers, Dataset Diatheca and Dataset Private Correspondence of the 20th and 21st centuries. **H_2.2, H_2.3, H_2.4.**

ACTIVITIES:

- upgraded Digitalia MUNI ARTS repository with an extended user community;
- access to datasets enriched with the use of CIDOC-CRM ontology;
- FAIRified datasets in the Digitalia MUNI ARTS repository will be made accessible;
- methodological and support activities related to the operation of the Digitalia MUNI ARTS repository and the implementation of Open Science in the humanities and arts.

OVERVIEW OF SUB-ACTIVITY OUTPUT CODES:[127]

H_2	Digitalia MUNI ARTS repository
H_2.1	CIDOC-CRM ontology metadata model
H_2.2	FAIRified dataset: Dictionary of Czech Philosophers
H_2.3	FAIRified dataset: Diatheca
H_2.4	FAIRified dataset: Private Correspondence of the 20th and 21st Centuries.

10.6.3. SUB-ACTIVITY 7.3 – UPGRADE OF THE ARCHAEOVAULT REPOSITORY AND FAIRIFICATION OF RELATED DATA

The ArchaeoVault repository is designed specifically for archaeological data. Due to the highly interdisciplinary nature of modern archaeological research, which closely combines the findings of the humanities and natural sciences (e.g. dating of individual preserved artefacts using advanced physical methods – isotope measurements, dendrochronological analyses, etc.), the existence of a separate archaeological repository appears highly functional and desirable. The ArchaeoVault repository is being prepared in close cooperation with other domain workplaces, in particular the Institute of Archaeology of the Czech Academy of Sciences Prague and Brno, which jointly operate the large research infrastructure Archaeological Information System (as such workplaces must do by law, its own

[127] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

repository, AIS, collects archaeological research reports; due to this specific purpose, it lies outside the scope of the OS II project). The preparation of the ArchaeoVault repository involves a significant proportion of methodological work: in cooperation with the professional community, a research report/analysis will be gradually developed and prepared, identifying best practices in the processing and digitization of archaeological data, and two metadata models for describing archaeological data. A methodology for processing archaeological data for the needs of the ArchaeoVault repository will be created based on these. The practical process of processing archaeological data will then be comprehensively described in the form of sample workflows which will be accessible to the user community. In practice, the entire procedure will be demonstrated on the RES-HUM dataset, which will also become a model example for the creation of future content for the ArchaeoVault repository. The upgrade of the ArchaeoVault repository will thus take place in close interaction with key domain institutions, taking into account all fundamental aspects of open science and FAIR principles and will demonstrate the possibilities for their use in curatorial and research practice. The sub-activity will include methodological activities related to the operation of the ArchaeoVault repository; Open Science will also be implemented in the humanities and arts.

Upgrade of the repository

In accordance with the Call, in the course of the project the repository will be (a) integrated with the NMD, to which mandatory metadata will also be provided, (b) connected to the EOSC common authentication and authorization infrastructure, (c) connected to the technical data transfer tools, methodologies and standards created by the EOSC-CZ IPs and the project ensuring the construction of NRPs, (d) registered in the NQF. Depending on the possibilities and domain practices, the repository will also be integrated into the international research environment. **H_3**

Development of metadata models

Both proposed metadata models anticipate the identification of individual data objects and their subsequent analysis. Based on this, a standardized metadata structure will be developed and iteratively tested, then subsequently documented and presented to the public. **H_3.1, H_3.2**

Preparation of the methodology

The project will develop a methodology for processing archaeological data. In the first phase, data and metadata requirements and their typology will be defined. The methodology will be published following the incorporation of the technical requirements of the repository. **H_3.3**

Preparation of the analysis

Identification of best practices in the processing and digitization of archaeological data. In the initial stage of work on the project, an analysis of best practices in the processing of archaeological data will be prepared. The curatorial practices of individual workplaces will be collected and subsequently analysed. Standardized procedures with the potential for universal application will be proposed based on the comparison of these practices. **H_3.4**

FAIRification of datasets

The RES-HUM dataset will undergo storage format optimization during the project solution; it will be cleaned and linked to related entities. Subsequently, the data will be checked for consistency and any errors will be eliminated. The resulting, rather extensive dataset will be published in the repository. **H_3.5**

Preparation of workflows
Workflows for data processing will be designed based on the previous outputs. These include the definition of input functionalities, the development and testing of related software tools, and their iterative testing and optimization. Finalized workflows will be documented and subsequently published for use. **H_3.6**

ACTIVITIES:
- upgrade of the ArchaeoVault repository, its extension to the community of archaeologists outside the Faculty of Arts of Masaryk University;
- development of metadata models;
- development of a methodology for processing archaeological data;
- analysis of best practices for the processing and digitization of archaeological data;
- FAIRification of the RES-HUM dataset and its publication in the ArchaeoVault repository;
- workflows made available for archaeological data;
- methodological and support activities related to the operation of the ArchaeoVault repository and the implementation of Open Science in the humanities and arts.

OVERVIEW OF SUB-ACTIVITY OUTPUT CODES128

H_3	ArchaeoVault repository
H_3.1	Archaeological metadata models
H_3.2	Archaeological Ontology metadata model
H_3.3	Methodology of archaeological data processing
H_3.4	Analysis: Identification of best practices in the processing and digitization of archaeological data
H_3.5	FAIRified RES-HUM dataset
H_3.6	Workflows for archaeological data

10.6.4. SUB-ACTIVITY 7.4 – CREATION OF A REPOSITORY FOR BIBLIOGRAPHIC DATA AND FAIRIFICATION OF RELATED DATA

During the project, the Czech Literary Bibliography (CLB) research infrastructure will prepare a Repository for Bibliographical Data as an important type of research data, and not only within the HUMA cluster. This repository will be created either as a separate instance of one of the software platforms supported by the NRP, or as a separate sub-community/collection within one of them. This decision will derive from the results of the internal analysis, which will include the first year of the project.

At the same time, the CLB will prepare two methodologies for the FAIRification of basic types of bibliographic and related data, namely: (a) a methodology for editing bibliographic data existing in analogue form (printed bibliographies, card indexes) and (b) a similar methodology for processing data from biographical dictionaries (primarily review sections). These methodologies will be verified on selected source data, which will then be made available in the form of three pilot datasets, on which the possibilities of converting originally analogue data into a structured database form will

[128] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

be tested. Key data sources with a significant overlap in other disciplines within the cluster were pre-selected as pilot examples: Retrospective Bibliography of Czech Literature (the largest bibliography in the domain in the form of a card index in the Czech Republic), Lexicon of Czech Literature (a key work of domestic humanities lexicography) and the Bibliographic Catalogue of the CS(S)R. Articles as a central national source of bibliographic data of a given type for the second half of the 20th century. Test datasets created in the course of work on the project will then be made publicly available and, in addition to them, a FAIRified dataset of the database of literary prices will also be created. Based on both methodologies, separate workflows will then be prepared for the needs of the SSH Open Marketplace as a central cluster platform for the promotion of open science at the European level and will also be available through this platform in English for an international audience. In order to ensure standardization in the field of literary bibliography, two metadata models will be created, namely a metadata model for the factual description of literary data and a model for the description and classification of literary works. The outputs of this sub-activity will then be presented in the form of educational material (online tutorials for working with the newly created repository, etc.).

For the Repository for Bibliographical Data itself, basic curatorial settings will be prepared and operational policies will be defined that will be in accordance with FAIR principles and the requirements of Open Science, while taking into account the specific parameters of bibliographic data (versioning, persistent identifiers, linking of bibliographic and authoritative data). In the future, the repository will be conceived as an open repository and will be offered for use to other bibliographic workplaces. Thanks to the inclusion of datasets from the European Literary Bibliography project, which is jointly developed by the Czech and Polish Literary Bibliographies and whose content is the extraction, harmonization and subsequent joint presentation of bibliographic data for individual national literatures (currently Czech, Polish, Finnish and Spanish data), the overlap of the repository for an international audience will also be ensured.

A total of six trips abroad are planned for the quality dissemination of the outputs to an international audience. The content of these will primarily be participation in prestigious domain events such as annual conferences of the EOSC initiative, the Association of Digital Humanities Organizations (ADHO) or DARIAH Annual Events, and related presentations of the project outputs. The sub-activity will include methodological activities related to the operation of the Repository for Bibliographical Data; Open Science will also be implemented in the humanities and arts.

Establishment of the Repository

In the first, analytical phase of preparation of the Repository for Bibliographical Data, the requirements of the user community for a specific bibliographic repository (metadata schema, user settings, etc.) will be analysed and the possibilities of existing software solutions will be compared. A specific technological solution will then be selected on the basis of these works. In the second, construction phase, the selected solution will be ready for sharp deployment: the repository will be connected to the NMD and registered in the NCR; the appropriate authentication and authorization infrastructure will be set up and the technical tools, methodologies and standards of the EOSC CZ platform will be implemented. **H_4**

Development of metadata models

The project will create previously non-existent metadata schemes for the needs of literary science, specifically for the factual (thematic) description and classification of forms and genres. A metadata model for the description of literary works, including standards for the description of mutual links and dependencies between them, will be created in parallel to this. **H_4.1, H_4.2**

Preparation of the methodology

As part of the project, methodologies for converting analogue data into a structured form will be prepared. The first will be devoted to bibliographic data and the second to biographical data. Both methodologies will be iteratively tested on pilot datasets during the project (cf. the FAIRification of datasets below) so that the proposed methodological procedures take into account model situations based on real practice. **H_4.3, H_4.4**

Preparation of materials for education

As part of the project, educational materials for working with the Repository for Bibliographical Data and that will practically outline the possibilities of its use to users will be prepared. **H_4.10**

FAIRification of datasets

The project will include the preparation of a total of four datasets. The first three will be designed as a pilot and will be designed primarily to test and verify the functionality of the aforementioned methodologies. In addition to these, the Literary Awards Database, which today no longer meets the standards of modern information science, will be completely cleaned, converted into a standardized format and enriched with a set of persistent identifiers. **H_4.5, H_4.6, H_4.7, H_4.8**

Preparation of workflows

Based on both proposed methodologies, related workflows will be created and published in English on the SSH Open Marketplace platform, which is the central European open science platform for the HUMA cluster. The availability of these workflows in English will ensure the future use of the project outputs for international audiences. **H_4.9**

ACTIVITIES:

- creation of a Repository for Bibliographical Data that is ready for live deployment;
- development of metadata models;
- preparation of methodologies for converting analogue data into a structured form;
- creation of educational materials for work with the Repository for Bibliographical Data;
- FAIRification of a set of datasets available in the Repository for Bibliographical Data;
- the relevant workflows will be made available on the SSH Open Marketplace platform;
- methodological and support activities related to the operation of the Repository for Bibliographical Data and the implementation of Open Science in the humanities and arts.

OVERVIEW OF SUB-ACTIVITY OUTPUT CODES[129]

H_4	Repository for Bibliographical Data
H_4.1	Metadata model for literary science
H_4.2	Metadata model for describing literary works
H_4.3	Methodology for converting domain bibliographies in analogue form into a standardized and structured FAIR form using AI tools
H_4.4	Methodology for converting professional lexicographical works into a structured and computer-readable form
H_4.5	FAIRified dataset Retrospective Bibliography of Czech Literature in conversion using AI tools
H_4.6	FAIRified dataset Pilot conversion of professional bibliographies in the form of printed bibliographies into a structured form using AI tools
H_4.7	FAIRified dataset Pilot conversion of professional lexicographic work into a structured and computer-readable form based on the example of the Lexicon of Czech Literature
H_4.8	FAIRified dataset Literary Awards Database
H_4.9	Making bibliographic methodologies available in the form of workflows on the SSH Open Marketplace platform
H_4.10	Educational materials for working with the Repository for Bibliographical Data

10.6.5. SUB-ACTIVITY 7.5 – DEVELOPMENT AND PILOT IMPLEMENTATION OF SUPERSTRUCTURE TOOLS FOR TKA HUMA REPOSITORIES

The sub-activity will focus on the technological development and implementation of tools for the development of the aforementioned repositories in order to achieve higher user comfort and improve the standard of services for end users. If possible, individual outputs will be developed with a more universal overlap, including for user communities outside the institution's own HUMA cluster. In the OS II project, we plan to develop software tools and related workflows for the incorporation of virtual reality functions as a key output of the "Set of superstructure tools for the development of TKA HUMA repositories".

The CNC will develop software for visualization of geolocation data in this KA. While this will be adapted to the needs of the DSpace repository system and harmonized primarily with linguistic or broader humanities data, it will be composed with a view to possible use outside these areas. Within this KA, UWB will prepare two software tools: software for the machine processing of audio data will include, inter alia, the development of modules for automatic speech recognition, diarization and possibly other related text processing tools, and a software tool for harvesting web documents will be created in addition to it. As part of the sub-activity, methodological activities related to the development of superstructure tools for humanities repositories and the implementation of Open Science in the humanities and arts will also be carried out.

Workflows for the VRE functions of CU ARTS repositories will be built primarily over archaeological collections. The prepared solution should include, among others, online visualization, 3D modelling or GIS functions, including the possibility of storing the results of these operations directly in the repository.

[129] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

As part of the project, three software tools will be developed for application in repositories of the HUMA domain cluster. The first one will be a tool for the visualization of geolocation data which will be primarily developed for the DSpace repository system. At the same time, two tools for processing language data will be developed: the first will be an online tool for the recognition of speech and its subsequent conversion into text, available via a web interface, and the second will be software for the automatic harvesting of texts from the web, especially for the needs of linguistic research. **H_5.1, H_5.2, H_5.3**

As part of the project, advanced virtual reality functions (e.g. for visualization, working with a map or 3D objects) will be implemented for the ArchaeoVault repository. As these tools will be directly integrated into the source repository, individual users will not have to install them separately, achieving improved comfort for users when working with the repository. **H_5.4**

ACTIVITIES:

- software development;
- set of documented software tools available on a public repository;
- preparation of workflows;
- methodological and support activities related to the preparation of superstructure tools for humanities repositories and the implementation of Open Science in the humanities and arts.

OVERVIEW OF SUB-ACTIVITY OUTPUT CODES[130]

H_5	Set of superstructure tools for the development of TKA HUMA repositories
H_5.1	Software for the visualization of geolocation data in DSpace repositories
H_5.2	Software for speech recognition via web interface
H_5.3	Web-based text harvesting software
H_5.4	Workflow for VRE function of ArchaeoVault pilot repository

10.7. KA 8 – THEMATIC CLUSTER ENVIRONMENTAL SCIENCES

PARTICIPATING PARTNERS: MU (TKA Guarantor), IOCB CAS, CU

Several needs of the research community have been identified for new repositories with environmental themes.

there are three new repositories in the area of chemical risks. In recent years, targeted analyses of selected chemicals have been replaced or supplemented by non-targeted analyses that store information on all (endogenous and exogenous) chemicals. In this case, however, extensive primary data from high-resolution mass spectrometry (repository for data from non-targeted metabolomics and exposomics) must be stored. In order to analyse chemical risks, it is necessary to supplement the chemical exposure data with information on chemical hazards, i.e. toxicological and ecotoxicological data. And if we want to study the impact of a contaminated environment on a specific population, it is necessary to link information about environmental contamination in different places by geocoding all information. This approach will also allow us to link environmental data with data on health and other factors that may affect humans (e.g. economic or social deprivation).

[130] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

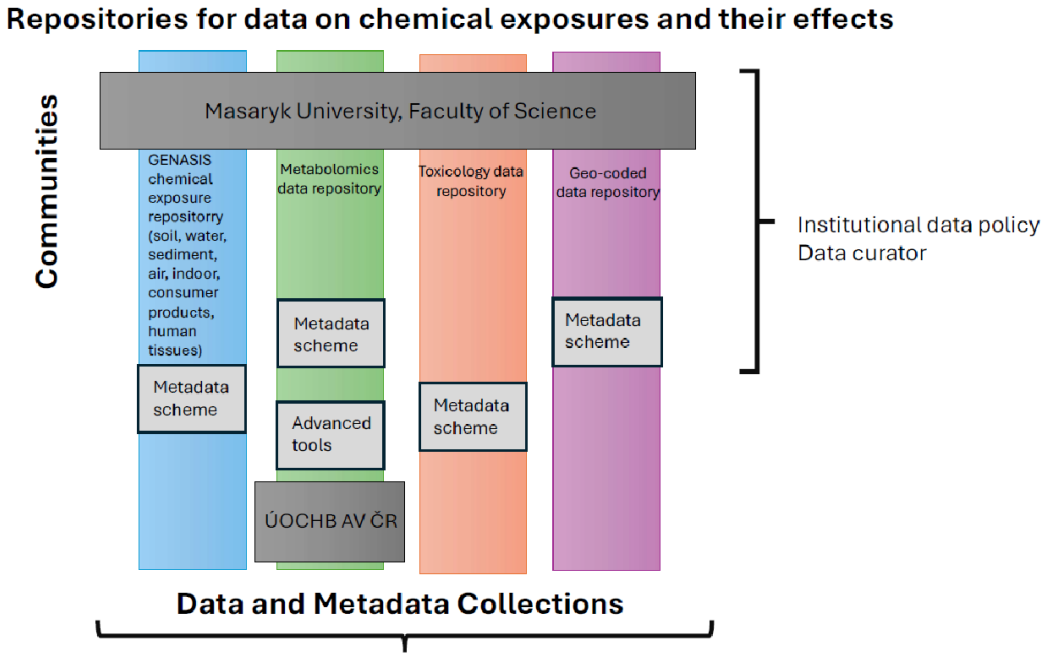

Fig. 10: Graphical diagram of sub-activities 8.1–8.4

However, new approaches to studying the environment are even more complex. The so-called triple environmental crisis we are facing involves not only environmental chemistry, but also climate change and biodiversity loss, all of which are closely related. Therefore, when creating new repositories, we will also focus on new repositories for zoological collections, genetic samples and reference image data. Scientific collections of zoological nature are defined as collections of alcoholic, dermoplastic and osteological preparations of various vertebrate and invertebrate species necessary for the interpretation and reproducibility of scientific data in the field of taxonomy and diversity. Repositories for genetic biomonitoring will enable the effective sharing of biodiversity data from the sequencing of mixed environmental samples and genetic data of wildlife, evaluation of the occurrence of taxa and their comparison with reference data. Reference image data are essential for basic research in the fields of botany, lichenology, paleoecology, for the protection of biodiversity, and for the training of classification models. By creating appropriate metadata models, we will ensure not only data interoperability within the cluster of repositories for biodiversity data, but also their connection to the international level (ESFRI infrastructure DiSSCo.cz or the GBIF biodiversity data aggregator).

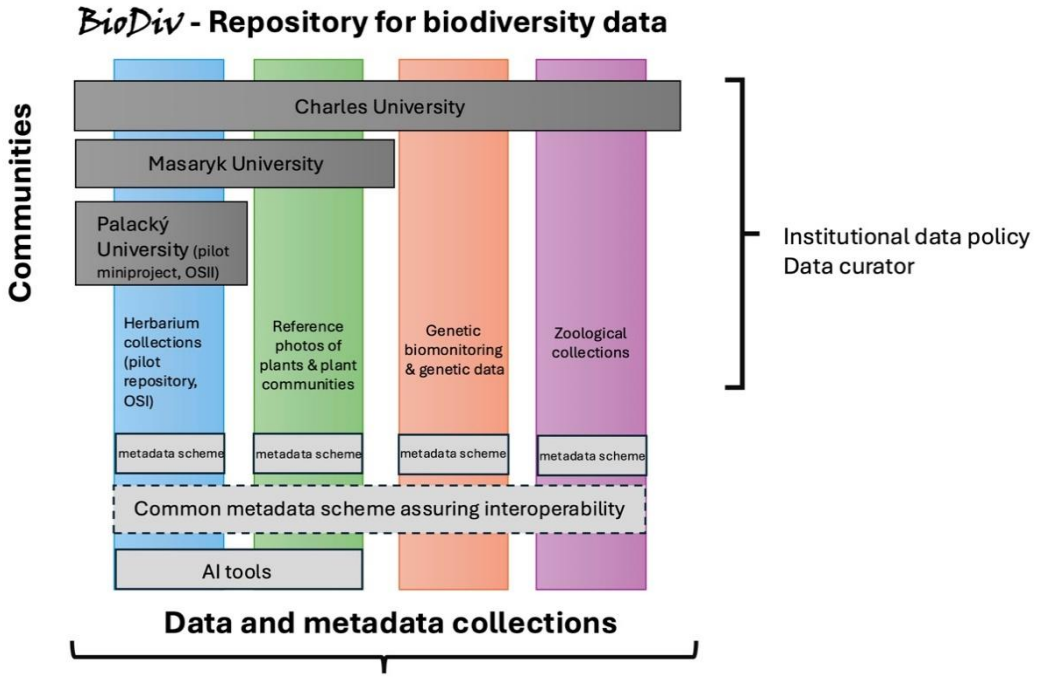

Fig. 11: Graphical diagram of sub-activities 8.5–8.7

In the key activity, we will focus on the:

- creation of metadata models, standards and methodologies enabling the processing of mass spectroscopic data, photographic records for the purpose of studying biodiversity, the genetic bank of wildlife, zoological collections, toxicological and ecotoxicological data, and GIS data.
- implementation of metadata standards to ensure semantic interoperability (interconnection with controlled dictionaries, thesauri and ontologies);
- assigning citable persistent identifiers to data objects or extracts;
- preparation of an application interface for the metadata exchange of managed datasets (DCAT) and an interface for data exchange;
- preparation of supporting materials, including instructions for the creation and validation of FAIR-ified datasets;
- analysis of other needs of the research community.

In all cases, the issues of ensuring the quality and interoperability of repositories and data in an international context, licensing and legal models, and connections to external data sources will be addressed.

The main benefits will be:

- standardized metadata models that will contribute to the consolidation of research domains and user communities that have so far operated without properly managed repositories;
- increasing the availability of existing domain-specific data (including their metadata measures);
- motivating the research community to participate in the EOSC and mobilise data stored outside public repositories;
- improved cooperation between institutions and teams at national level in the areas concerned;
- visibility of Czech research and its results, stronger involvement in international projects and partnerships;
- strengthening the position of the Czech Republic as a potential host country of EIRENE RI.

10.7.1. SUB-ACTIVITY 8.1 – UPGRADE AND EXPANSION OF THE GENASIS REPOSITORY FOR THE PRESENTATION OF DATA ON THE CHEMICAL CONTAMINATION OF INDIVIDUAL ENVIRONMENTAL MATRICES AND CHEMICAL EXPOSURE OF THE POPULACE[131]

The GENASIS environmental information system has[132] been gradually developed in the last two decades at Masaryk University's RECETOX centre in response to the need for a repository that would jointly present research data on the chemical contamination of individual components of the environment (open air, precipitation, surface water, sediment, soil, etc.). However, the availability of such data is essential not only for environmental research, but also for the development of international, national and regional policies (for example, the implementation of international agreements on the protection of the environment and human health from the effects of chemicals). Therefore, the system has been gradually expanded to include both additional matrices (such as indoor air or dust) and newly monitored emergent substances, and its use was also offered to other research institutions producing data on environmental contamination as part of joint (and international) projects. At the international level, the GENASIS metadata model was subsequently used by the United Nations Environment Programme (UNEP) for the preparation and implementation of the Global Monitoring Plan of the Stockholm Convention[133]. In this case, GENASIS serves as the primary storage area for individual data points for subsequent aggregation and reporting.

The main added value of GENASIS is its function as a single database of harmonized data from multiple providers. When entering data into the database, their control, cleaning and possible transformation into a harmonised format takes place either on the part of the data provider or on the part of the GENASIS data manager. Unification into a harmonised structure ensures the comparability of data across the database, regardless of their source. This will rapidly increase their usability for other applications, without transferring this work to end users of data, who will always be limited in their possibilities and knowledge of how to perform such harmonization.

A specific feature of GENASIS is that it is not intended solely for researchers, but for a broad range of users, including public administration, the commercial sector and the general public. For these purposes, tools for the visualization, analysis and interpretation of data, including map outputs, statistical modules or analysis of long-term trends, as well as tools for data aggregation and their transfer to international systems, were gradually built over the GENASIS database. For this reason, it is advisable to maintain the entire system on the existing repository, but while investing in updating it so that it fully meets the current needs of the professional community and FAIR principles, as well as resolving the issue of licensing and legal regimes.

The modernization of the GENASIS repository planned under the OS II project includes its extension towards new types of matrices (e.g. human tissues such as blood, urine or breast milk) as well as newly monitored substances (phthalates, bisphenols, UV filters, mycotoxins, etc.). The problem of chemical risks is currently addressed at the European level by the HEU Partnership for assessment of risks from chemicals (PARC), in which RECETOX is an active participant. Therefore, when creating ontologies and metadata models of the upgraded repository, emphasis will be placed on their full compatibility with the standards created in PARC. To do this, it will be necessary to modify the data structures and descriptive metadata and extend the metadata model with the attributes required in the basic NMD metadata model. To ensure flexibility in conversions to different output formats, the data model

[131] In the description of the KA, the planned activities are linked to the implemented outputs. The outputs are marked with a code. For TKA ENVIRO, this is in the format **E_x** (key activity), **E_x.x** (partial, i.e. part of the key activity).
[132] www.genasis.cz.
[133] www.pops-gmp.org.

will be documented in machine- and human-readable formats and an API will be constructed for transferring metadata to the NMD. The introduction of semantic interoperability, i.e. linking metadata to controlled dictionaries and ontologies, will be essential. **E_1.1**

To ensure citability, the next crucial step will be to assign persistent identifiers to data objects or extracts. Globally recognized persistent identifiers will be used for important entities such as CAS and InChI for chemicals, IGSN ID for samples or RAiD for projects and studies. The problem with this is that it is difficult to cite a database that is undergoing continuous updates and data increments. However, individual data records can be considered immutable (changes are only possible if an error is found), as well as imported datasets. Persistent identifiers can also be assigned to exported extracts from the repository. It will thus be possible to regularly publish steady-state datasets, including continuous updates and data increments (1-4 times a year depending on the nature and frequency of the input data update). These steady-state machine-generated datasets/extracts will be automatically published in the NMD. However, it will also be possible to create extracts based on individual requests for data (thematically published datasets will be created according to typically used criteria).

As part of the upgrade of the GENASIS repository, we also plan to change the data release mechanism so that all data extracts are registered in the open access management system for the EIRENE/RECETOXl infrastructure. In all cases, data redistribution will be treated with respect to individual combinations of license terms. The licensing mode will then create clear rules for working with data of external providers for all users of the system (this is especially important due to the fact that data of external providers, including foreign institutions, may be transferred to other programs or projects). **E_1**

In the first phase, in accordance with the new metadata model, the existing data of the coordinating institution will be harmonized and provided with metadata, and in subsequent stages, the data of other institutions will be connected via "mini-projects". All data files will be supplemented with the appropriate metadata sets. We anticipate that in the future the upgraded repository and all its tools will be used internationally (both as part of the ESFRI data services of the EIRENE research infrastructure and as a mechanism for collecting and presenting the data of the Global Monitoring Plan), which will support the long-term sustainability of the upgraded GENASIS repository. **E_1.2**

ACTIVITIES:

- the extension of the GENASIS repository for the presentation of data on the chemical contamination of environmental matrices entails the reprogramming of the structure of the upgraded repository and its connection to the NRP; **E_1**
- extension of the metadata model of the GENASIS chemical exposure data repository, which will allow the storage of data from new matrices and emerging pollutants in accordance with the standards defined
- in the PARC partnership; **E_1.1**
- creation of metadata sets accompanying all data on chemical pollution of environmental matrices in the GENASIS repository; **E_1.2**
- creation of harmonised datasets on the chemical contamination of environmental matrices in accordance with the newly defined metadata model of the GENASIS repository; **E_1.3**
- preparation of supporting materials (standard operating procedures and tutorials) for the creation and validation of FAIRified datasets. **E_1.4**

OVERVIEW OF SUB-ACTIVITY OUTPUT CODES[134]

E_1	Upgraded GENASIS repository
E_1.1	Innovative metadata model of the GENASIS chemical exposure data repository, which will allow the storage of data from both new matrices and emerging pollutants
E_1.2	Metadata sets in the GENASIS repository accompanying all data on the chemical contamination of environmental matrices
E_1.3	FAIRified data sets in the GENASIS repository on the chemical contamination of environmental matrices
E_1.4	Standard operating procedures and tutorials for the creation and validation of FAIRified datasets

10.7.2. SUB-ACTIVITY 8.2 – REPOSITORY FOR THE STORAGE OF DATA FROM NON-TARGETED MASS SPECTROSCOPIC ANALYSES FOR HUMAN EXPOSURE ASSESSMENT

For decades, environmental analytical chemistry has focused on the targeted analysis of several selected groups of substances with demonstrably harmful consequences (the results of these targeted analyses are then presented, for example, in the aforementioned GENASIS repository). However, as the number of anthropogenic substances used in industrial or agricultural production grew and the volume of toxicological information about them expanded, the list of chemicals that need to be monitored in the environment or human matrices also expanded. At the same time, it proved impractical to focus only on substances for which we have currently available information, and methods of broad non-targeted screening began to develop. These methods collect information on the presence of the entire spectrum of exogenous and endogenous substances in environmental matrices or human tissues, regardless of whether these are substances that we can currently identify. This is therefore a kind of digital biobanking: we generate a complete record of the signals of all substances present; these must then be retained for a long period of time, while facilitating repeated searches by various users. Unlike targeted analyses, which do not require the storage of primary data, but only of the resulting measured concentrations (see GENASIS), in this case primary records, to which various methods of analysis will continue to be applied, really are required, while all intermediate products of these analyses must be retained for further use.

The resulting repository is actually a fundamental functional extension of the GENASIS repository, so it, like all other repositories related to environmental and human exposure to toxic substances, needs to be built as a single complex so that all data can be linked for the purpose of analysing environmental and health risks. To illustrate the whole cascade: Primary spectroscopic data represent a huge amount of data, comparable to genomic data, and, like untargeted genomic data, are a resource for long-term research. The product of primary analyses are tables of individual signals for future potential identification. The next step is qualitative, i.e. determining the chemical identity of the signal, and the last step is quantitative, i.e. determining the concentration of a given substance. While only this result can be saved in the GENASIS repository, all others must be available to users elsewhere. In the future, the entire complex of these repositories is intended to become the core of data sources for the EIRENE research infrastructure.

[134] The sub-activity output codes are linked to Annex 3 Schedule of Key Activities and chap. 11 Main planned outputs/products of Feasibility Study.

A specific feature is that it is not a static repository, but a source for continuous further mining, which requires advanced bioinformatics techniques. These cascades of bioinformatics tools also need to be constantly developed so that the proportion of correctly identified or quantified signals gradually increases. This key activity therefore includes the development of these tools, as will be described below. As many teams from across the globe will be involved in the development of this software, the tools will have to be implemented in a shared environment – in this case, the Galaxy environment administered by the ELIXIR infrastructure. The interconnection of the newly constructed EIRENE and ELIXIR data and computing capacities will offer the research community resources that were previously unavailable.

For non-targeted analyses of human samples, it is an indisputable advantage that anthropogenic pollutants and products of metabolism can be measured using the same method (and thus combine the study of human exposure/exposome and metabolome) to add additional layers to multiomic analyses. However, this necessitates the storage of large volumes of data from high-resolution mass spectrometry (primary and derived) in the long term and the facilitation of not only repeated searching, as the volume of our knowledge about the exposure of the population to harmful substances and their harmful effects increases, but also their connection with data characterizing the human genome, transcriptome, proteome or microbiome. This is the only way of working towards the discovery of new biomarkers of the effect of chemical exposures or biomarkers of the development of chronic conditions and make significant progress in personalized medicine and prevention.

While a number of repositories for storing data from non-targeted metabolomics exist across the world, their metadata structure does not allow for the sorting of available data and repeated analysis of their relevant subgroups. Therefore, in this activity we focus on creating a new metadata model in which the minimum set of metadata will be significantly expanded. The goal is to create a structure that will open up the possibility of integrating large data from different domains and analysing the genetic and non-genetic factors behind the development of chronic diseases in the future. At the same time, it is not our ambition to build a global repository, but only a generally accepted model of such a repository; this will be used primarily by the EIRENE European infrastructure reference laboratory, or by a cluster of reference laboratories. In the next steps (outside of this project), we envisage close cooperation between the ESFRI infrastructure EIRENE and EMBL/EBI to create a mirror large-capacity repository that is managed by EMBL/EBI and used internationally. This process will ensure international interoperability and the possibility of the integrated mining of metabolomic data. However, in order to be able to connect these data with data from genomics and other "omics" in the future, it will be necessary to establish closer cooperation with these professional communities and jointly build data interfaces and paths to connect relevant datasets. However, this goes beyond OS II.

As part of the OS II project, we will focus on the creation of a repository that will allow the storage of all stages of metabolomic data (from primary records from gas or liquid chromatography and high-resolution mass spectrometry through annotations, identification to quantification) as well as extensive metadata enabling not only their repeated analysis, but also future links to data from other domains.

Based on our experience with long-term storage and processing of these data, a formal data model will be created in connection with the ontology used and metadata standards will be introduced and published. Semantic interoperability will be ensured by interconnection to controlled dictionaries, thesauri and ontologies (these are essential in particular for chemical classifications, e.g. CAS, InChI). Individual data objects will be assigned citable persistent identifiers (in this case, the primary data records are invariable and each newly created set of secondary data will be provided with a new identifier) and application interfaces will be created for the exchange of metadata of managed datasets

and for the exchange of data itself. Persistent identifiers will also be assigned to exported outputs from the repository (i.e. secondary data, e.g. annotated signal tables) to ensure citability. The licensing model for sharing and using primary and secondary data will be resolved, also taking into account the sensitivity of potential interconnection with data from other domains. The new repository will be registered in the NCR, mandatory metadata about the datasets managed in this repository will be provided to the NMD. User authentication and authorization processes will utilise the common authentication and authorization infrastructure ("AAI") operated by the NRP and, wherever possible, will be connected to technical data transfer tools, methodologies and standards of the NRP.

In the first step, the repository will be filled with available data and metadata of the EIRENE/RECETOX infrastructure, optimized and validated, and in the next step it will be offered to the broader community for use. In connection with this activity, one trip to a metabolomics conference abroad is planned for the sharing of experience with the international harmonization of standards and discussion with the professional community.

OVERVIEW OF SUB-ACTIVITY OUTPUT CODES[135]

E_2	Repository for storing data from non-targeted mass spectroscopic analyses for human exposome assessment
E_2.1	Metadata model for primary mass spectroscopic data from non-targeted metabolomics and exposureomics of human samples
E_2.2	Metadata sets relevant to primary mass spectroscopic datasets from non-targeted analyses for exposure and metabolome studies
E_2.3	FAIRified datasets of primary mass spectroscopic data from non-targeted analyses for exposure and metabolome studies

10.7.3. SUB-ACTIVITY 8.3 – REPOSITORY FOR STORING TOXICOLOGICAL AND ECOTOXICOLOGICAL DATA

The analysis of environmental and health risks associated with environmental contamination by chemicals requires the supplementing of data on the occurrence of these substances in environmental matrices and human tissues (stored in the GENASIS repository) with data on their harmful effects (hazards), i.e. toxicological and ecotoxicological data. However, no uniform repository of such data exists worldwide and toxicological data are published only in scientific publications.

The matter is further complicated by the fact that living organisms are never affected by one substance alone (as is commonly tested for in toxicity tests), but by a complex mixture. Therefore, when analysing chemical risks, we should ideally take into account the simultaneous synergistic or antagonistic action of all substances present. One way to approach this ideal is the concept of so-called "adverse outcome pathways" (AOP). In this case, the individual biological pathways on which toxic substances may act are monitored and information on the strength of the effect of all relevant stressors is included. However, the whole concept runs up against the problem of the lack or poor availability of toxicological data.

[135] The sub-activity output codes are linked to Annex 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

We will carry out an analysis of the needs of the international research community in cooperation with the European Partnership for Chemical Risk Analysis (PARC), which brings together leading European toxicologists. This will be followed by the design of the structure of the toxicological repository and the metadata model; again, both concepts will be discussed with the international scientific community. International consensus and interoperability are crucial for the future success of the project, as the repository will, like the two previous ones, be part of the European ESFRI EIRENE infrastructure and will be used internationally.

Following the creation of a formal data model, metadata standards will be introduced and published. As in the case of the GENASIS exposure repository, semantic interoperability will be ensured by interconnection with controlled dictionaries, thesauri and ontologies. The linking of exposure data (GENASIS) and hazard data (this repository) will be possible thanks to a uniform chemical classification (CAS, InChI). In order to ensure citability, persistent identifiers, including continuous updates and data increments, will be assigned to regularly published extracts from the repository. These established machine-generated datasets will be automatically published to the NMD. However, it will also be possible to create extracts based on individual data release requests. The new repository will be registered in the NCR and mandatory metadata to the NMD will be provided about the datasets managed in this repository. A licensing model will be set up and user authentication and authorization will utilise a common authentication and authorization infrastructure. The repository will be pilot tested on the available toxicological and ecotoxicological data and metadata of the RECETOX centre before being opened to the international scientific community (through both “mini-projects” and the EIRENE infrastructure).

OVERVIEW OF SUB-ACTIVITY OUTPUT CODES[136]

E_3	Repository for storing toxicological and ecotoxicological data
E_3.1	Analysis of the needs of the research community in the field of toxicological and ecotoxicological data management and available solutions for the preparation of a new repository and its metadata model
E_3.2	Metadata model for toxicological and ecotoxicological data
E_3.3	Metadata sets for toxicological and ecotoxicological data
E_3.4	FAIRified pilot datasets of toxicological and ecotoxicological data

10.7.4. SUB-ACTIVITY 8.4 – NEW REPOSITORY TO FACILITATE THE INTERCONNECTION OF GEOCODED DATA FROM DIFFERENT DOMAINS

If we are to study the impact of environmental factors on health, we need to link long-term data on human health (from long-term monitoring of population cohorts or patient health records) with information on all possible types of environmental stressors (chemical pollution of air, soil, and drinking water, natural disasters, heat, drought, reduced biodiversity) in the place where they live. At the same time, we can also monitor economic or social influences (e.g. socially excluded localities) in specific places. The current assessment of all factors affecting humans and their health is the principle of the human exposure concept. However, this holistic approach is transferable to the study of the health of any population (the so-called eco-exposome or the "one health" principle).

[136] The sub-activity output codes are linked to Annex 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

In order to monitor the health impacts (health data, results of biological material analysis) of all factors operating in one place at the same time, this information must be stored in relation to the geographic information system. The repository for geocoded data built as part of this project will enable the interconnection of GIS layers from different domains into a single functional system that will allow quantification of the environmental risks of individual areas.

Similar repositories are also being built in other places, for example as part of the European Human Exposure Network (EHEN), or the US NEXUS project. In this project, we will therefore proceed in close cooperation with international partners and ensure compliance with international standards, interoperability, and the long-term usability of data. In this case, however, the intention is to create a Czech model that collects a maximum of locally relevant data that cannot be taken from the international environment.

We will operate a map server with an application interface in the WMS/WFS standard. When describing the metadata profiles of each map layer, we follow the ČSN ISO 19107 standard (Geographic information – Spatial diagram). The analysis of the needs of the research community will be carried out in cooperation with the PARC, IHEN and NEXUS projects and will be the key to further specifying the structure of the repository, its metadata model and metadata standards. The new repository will be registered in the NCR and mandatory metadata about the individual mapping layers managed in this repository will be provided to the NMD. These data extracts will be provided with persistent identifiers to ensure citability. Authentication and authorization of users will be used by a common AAI operated by the NRP.

The repository will be piloted using long-term GIS layers available at Masaryk University. Other partners will be motivated to share new types of data, to cooperate across domains and to use available information in various research domains and application sectors. If the licensing terms allow, we will offer map layers to users for download, but in particular we will provide access to managed map layers for the user community through this map service. Furthermore, we will develop geoprocessing services that will provide information derived from individual map layers based on the entered spatial coordinates (this will allow to process the results even for individuals who do not control ArcGIS.

OVERVIEW OF SUB-ACTIVITY OUTPUT CODES[137]

E_4	A new repository for linking geocoded data from different domains
E_4.1	Analysis of the needs of the research community in the field of linking geocoding data from different areas and available solutions for the preparation of a new repository and its metadata model
E_4.2	Metadata model for geocoded data repository

10.7.5. SUB-ACTIVITY 8.5 - REPOSITORY FOR REFERENCE IMAGE DATA ON FLORA AND PLANT COMMUNITIES

Reference image data are photographs of objects (living plants, their communities or partial organs, such as pollen grains or microscopic sections of wood, etc.) that are classified clearly, accurately and systematically. These are essential as reference samples for basic research in the fields of botany, lichenology, paleoecology, but also for applied research and biodiversity protection. In areas where the machine identification of environmental samples is used to a greater extent

[137] The sub-activity output codes are linked to Annex 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

(such as the area of pollen analysis), their importance for the training of classification models is completely irreplaceable.

These data are already used by the scientific community both at home and abroad thanks to accessibility within applications such as the Czech botanical portal Pladias, the European vegetation portal[138][139], the Czech palynological database[140], the Neotoma[141] palaeoecological database, and the botanical photo gallery[142].

The partners of the OS II project (MU, CU) together with the Institute of Botany of the CAS, constitute the largest administrator of reference botanical data in the Czech Republic. The managed data is stored in fragmented form, i.e. by individual institutions and sub-projects from which the data was created, without proper integration, a unifying metadata description and unified data management compliant with FAIR principles. The transfer to the repository platform is already being successfully applied to scans of herbarium items as part of the pilot activities of the NRP project, and we will use the experience gained and the activated community for another type of data where both basic requirements for a repository are present: the need for large storage capacity (especially for machine-generated data from microscopic slides) and connectivity and citability during their publication – in addition to their use in research, reference data are also very necessary and widely used for the presentation of botanical taxa/data to the general public.

Therefore, we will create a separate Invenio repository for reference image data within the National Repository Platform with a suitable metadata model and that meets all requirements (documentation, policy, methodological support) for integration within and the movement of data to the NMD. At the level of data preparation for import into the repository, teams will first be mobilized within both institutions, then the data obtained will be standardized according to the metadata model and finally published within the repository.

The main outputs of the sub-activity will be: a functional repository with a national focus and sufficient storage capacity for reference botanical image data; a metadata model for the included data, its integration within a cluster of biodiversity repositories within the NRP with a view to interoperability, as well as within transnational activities (in particular ESFRI DiSSCo.eu), and other elements of repository operation. FAIRified datasets obtained by mobilizing the broad scientific community within the OS II project. The activity includes foreign trips to ensure the interoperability of repository platforms and metadata standards with the Distributed System of Scientific Collections[143] and JACQ.

OVERVIEW OF SUB-ACTIVITY OUTPUT CODES[144]

E_5	Repository for reference image data of living plants and plant communities
E_5.1	Metadata model for reference image data of plants and plant communities
E_5.2	Metadata sets for reference image data of plants and plant communities
E_5.3	FAIRified flora and vegetation image reference collection dataset

[138] https://pladias.cz.
[139] https://floraveg.eu.
[140] https://botany.natur.cuni.cz/palycz/.
[141] https://www.neotomadb.org/.
[142] https://www.botanickafotogalerie.cz.
[143] http://discco.eu.
[144] The sub-activity output codes are linked to Annex 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

E_5.4	Methodology for creating a reference collection of image data of flora and vegetation, creating data and metadata, and their FAIRification, storage and management

10.7.6. SUB-ACTIVITY 8.6 – REPOSITORY FOR GENETIC BIOMONITORING AND GENETIC DATA

Genetic biomonitoring enables the large-scale analysis and effective sharing of biodiversity data from environmental samples using sequencing of pooled samples and evaluation of the occurrence of taxa compared with reference data. In recent decades, genetic methods for monitoring wildlife have become widespread, yet they still lack a unified national platform for data storage. Therefore, we will create a new repository for data from genetic biomonitoring and genetic data of wildlife. The repository will be oriented across taxonomic groups and will include in particular the following types of data:

a/ image data of genetic biomonitoring of organisms, which are obtained by a robotic approach and are linked to the results of metabarcode sequencing of wild organisms obtained from environmental samples, such as mass catches of insects;

b/ data from the measurement of genome size and ploidy, possible other characteristics of the cell nucleus of wild organisms, obtained from flow cytometry, i.e. from an area where the Czech Republic is the world leader in the application of these methods to research on evolution and biodiversity;

c/ cytogenetic image data from fluorescence microscopy of mitotic and meiotic nuclei of cells. In this area, the Czech Republic is also one of the world leaders in the application of cytogenetic methods to research on the diversity and evolution of organisms. The aforementioned datasets are currently stored at several academic workplaces and museums in the Czech Republic – both at the project partners (Charles University, Masaryk University), but also at non-partner institutions (Institute of Botany of the CAS, Institute of Experimental Botany of the CAS, Biological Centre of the CAS, Institute of Animal Physiology and Genetics of the CAS, Institute of Vertebrate Biology, Natural History Museum of the National Museum, etc.) The main outputs of sub-activity E03 will be: a functional repository with a national focus and sufficient storage capacity for genetic biomonitoring and genetic data; its integration into a cluster of biodiversity repositories within the NRP with a view to interoperability, as well as within transnational activities (in particular ESFRI DiSSCo.eu); controlled access to and storage and management of data in a single location through AAI within NDI and other requirements for the operation of the NRP repository; uniform methodology and standards. Further substantial outputs will be the publication of FAIRified datasets, which will be obtained by mobilizing the broad scientific community within the institutions involved in the OS II project.

OVERVIEW OF SUB-ACTIVITY OUTPUT CODES[145]

E_6	New repository for the genetic biomonitoring and genetic data of wild organisms
E_6.1	Metadata model for genetic bank of wild organisms
E_6.2	Metadata sets for the genetic biomonitoring and genetic data of wild organisms
E_6.3	FAIRified dataset of genetic monitoring of wild organisms (plants and fungi)
E_6.4	Methodology for the creation and FAIRification of data and metadata for the genetic monitoring of wild organisms

[145] The sub-activity output codes are linked to Annex 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

10.7.7. SUB-ACTIVITY 8.7 – REPOSITORY FOR ZOOLOGICAL COLLECTIONS

Scientific collections are managed by academic (universities and institutes of the Academy of Sciences) and museum institutions (especially the Natural History Museum of the National Museum and the Moravian Museum, as well as also many regional museums). Scientific collections of a zoological nature are defined as collections of alcohol-preserved, dermoplastic and osteological preparations of various vertebrate and invertebrate species, which constitute the cornerstones for the interpretation and reproducibility of scientific data in the field of taxonomy and diversity at the national and transnational level. These are used as the material basis for the creation of digital descriptions (the "digital twins" concept) and image data (high-resolution photographs and scans, including 3D images) that allow them to be made available online. At the same time, they are used as a material for advanced imaging and analytical techniques (genetic, tomographic, etc.). At the same time, each collection item is also proof of a find at a given place and time, and is thus a source of data on the distribution of organisms.

The creation of the new repository is based on the needs of the research community in the Czech Republic, where the management of image data and metadata of zoological collections is currently highly fragmented, i.e. divided between individual institutions (in particular Charles University, the Institute of Vertebrate Biology, and the National Museum), is not properly integrated, and lacks a uniform metadata standard and uniform data management that would comply with FAIR principles. By creating the appropriate metadata model, we will ensure data interoperability both within the BioDiv cluster of biodiversity data repositories and at the international level – linking the DiSSCo.cz infrastructure to ESFRI and the world's largest biodiversity data aggregator, GBIF[146].

As part of the project, we will create a separate Invenio repository for reference image data from zoological collections within the National Repository Platform with a suitable metadata model and that meets all requirements (documentation, policy, methodological support) for integration within and the movement of data to the NMD. At the level of data preparation for import into the repository, the obtained data will be standardized according to the metadata model and finally published within the repository.

OVERVIEW OF SUB-ACTIVITY OUTPUT CODES[147]

E_7	New repository for zoological collections
E_7.1	Metadata model for zoological collections
E_7.2	Metadata sets of zoological collections
E_7.3	FAIRified dataset of zoological collections
E_7.4	Methodology for the creation of data and metadata of zoological collections, their FAIR-ification, storage and management

[146] https://www.gbif.org.

[147] The sub-activity output codes are linked to Annex 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

10.7.8. SUB-ACTIVITY 8.8 – DEVELOPMENT AND PILOT IMPLEMENTATION OF SERVICES FOR THE ANALYSIS OF ENVIRONMENTAL AND BIODIVERSITY DATA

As described above, the new repository of data from non-targeted, high-resolution spectroscopic analyses will be a valuable source of information that can be extracted by different users over a very long period of time. However, this is conditional on the existence of advanced software tools for searching such data, i.e. modules for signal cleaning, peak identification, prediction of mass spectra, and qualitative or quantitative analysis. This is not a single piece of software, but a rich cascade of tools that users can combine as needed. If such tools are of sufficient quality, documented, validated and openly available, it will contribute to the standardization of the entire data mining process and the credibility of the information obtained.

Therefore, following the newly built repository of metabolomic data, this project will build advanced tools for the analysis of data from non-targeted gas chromatography/mass spectroscopy (GC/MS) and liquid chromatography/mass spectroscopy (LC/MS) analyses for metabolomics and functional exposomics in order to improve the availability and reusability of research data according to the needs of the research community and the commercial sector. All these tools for working with available data will be implemented in cooperation with the ELIXIR infrastructure (and its Czech coordination node at IOCB CAS) in the Galaxy environment, which ensures their full documentation and open access for all users. This will contribute to reducing barriers to the interdisciplinary use of research data on a global scale and ensure interconnection with international activities and the standardization and wide availability of all tools. The built services will also include the preparation of new spectral libraries, standard operating procedures for data production and processing, and training materials and tutorials. In connection with this activity, one trip abroad to the Galaxy community conference is planned to share experience with the implementation of metabolomic tools and to discuss it with the professional community.

For the needs of biodiversity data repositories, we will develop and apply artificial intelligence software tools related to (i) image analysis of samples of a biodiversity nature, (ii) the application of Large Language Models (LLM), which will enable the automated reading of herbarium labels and records of zoological collections for the purpose of reliable, massive and effective access to metadata of biological collections to the general professional and lay public, and which will be part of biodiversity data repositories under the NRP.

ACTIVITIES:

- pipelines for the analysis of mass spectroscopic data from non-targeted metabolomic and exposure analyses; **E_8**
- AI tools for the image analysis and automatic reading of herbarium labels. **E_9**

OVERVIEW OF SUB-ACTIVITY OUTPUT CODES[148]

E_8	New bioinformatics tools for the analysis of mass spectroscopic data from non-targeted metabolomic and exposome analyses
E_9	AI tools for the image analysis of biological objects and the automatic reading of herbarium labels.

[148] The sub-activity output codes are linked to Annex 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

10. 8. KA 9 – THEMATIC CLUSTER SENSITIVE DATA

PARTICIPATING PARTNERS: CESNET (TKA Guarantor), MU, UPOL, UWB, SI CAS, JHIPC CAS

The Thematic Cluster Sensitive Data will add a layer of sensitive data management to the project across the NRP and NDI space. It will follow on from the support of activities essential to the FAIRification of data in repositories and work on licensing and other legal regimens/models (in the NRP currently for data in "non-sensitive", i.e. public mode). In the NRP and NDI, there is an infrastructure enabling data high-level management, but without the tools and functionalities necessary for the management of sensitive data. TKA SENSI will create detailed and instructional procedures for the FAIRification of sensitive data, tools or services applicable in the life cycle of sensitive data management, and last but not least, we will focus on the good practice of sharing and processing sensitive data (e.g. in Trusted Research Environments[149]) arising, among others, from cooperation between the academic and private spheres. Emerging or upgraded repositories, which will be integrated into the NRP under OS II and are intended for the storage of sensitive data, will develop these policies, procedures and services or tools in cooperation with TKA SENSI experts. This Thematic Key Activity aims to create clear and harmonised rules for accessing sensitive data that cannot be shared under current conditions. Inspired by the so-called FAIR principles,[150] we will create a methodology for the FAIRification of sensitive data applicable throughout the NDI and all repositories containing sensitive data. Following the technical tools being created in the NRP and NDI projects, such as the Fair Implementation Profile[151] (including FIP Wizzard[152]) or the training of data support staff and other experts within the community, our activities will focus on administrators and data curators of repositories, data providers, and their cooperation with data applicants. A methodology for the FAIRification of sensitive data will be created on the basis of cooperation with emerging repositories and the experience of experts from TKA SENSI.[153] Methodology and resulting rules for the management of sensitive data in the NRP, or NDIs will be the basis for training for academics managing sensitive data. Our TKA will prepare training materials for these courses, which will be led by our TKA experts.[154]

Within the project, seven repositories containing sensitive data will be created in individual TKAs. The TKA SENSI team will work closely with four of them and regularly hold consultations on the preparation of the FAIRification methodology for sensitive data, services and tools. We will work with the other three repositories as required and throughout the duration of the project.

These FAIR principles will be applied by the TKA team in the following activities across the NDI:

- **Traceability** – NRP repositories containing sensitive data will contain datasets of properly described metadata; the handling of metadata of sensitive datasets may vary across NRPs according to the needs of repositories; persistent identifiers and metadata are the basis for data traceability and support for the reproducibility of results obtained from the analysis of sensitive datasets.

[149] Secure platform for management and analysis of sensitive data, https://ukhealthdata.org/developing-technology-services/trusted-research-environments/.
[150] https://openscience.muni.cz/fair-a-open-data/fair-data-a-jak-na-ne/hlavni-principy-fair.
[151] https://www.go-fair.org/how-to-go-fair/fair-implementation-profile/.
[152] https://fip-wizard.readthedocs.io/en/latest/about/about.html.
[153] Cf. chap. 9 Organizational structure, project management and description of the roles of the implementation team.
[154] The training materials are an output of the OS II project. The training courses themselves will be realized in cooperation with the EOSC CZ Training Centre, which will provide adequate capacities, technical facilities and organizational support and are not included in the budget for OS II.

- **Accessibility** – transparent rules for access to data and clear conditions for the use of data defined according to the needs of each repository will be established; we will create a service of (to some extent automated) workflows enabling the analysis of sensitive data within the NDI without the need for access by the applicant or transmission and processing in a secure interactive computing environment.
- **Interoperability** – from the point of view of TKA SENSI, it is important that sensitive data in repositories are in standard and open formats that allow data to be connected across datasets, while at the same time establishing security rules for such connection or subsequent sharing.
- **Reusability** – we will build on the technical quality control in individual repositories and focus on the quality of data based on the control of the content of the dataset (known as plausibility); TKA SENSI will create transparent conditions under which sensitive data or access to them can be requested, including a machine-readable licence allowing automated systems to perform data operations in accordance with licensing conditions without the need for human intervention.

In addition to the general methodology for the FAIRification of sensitive data and establishment of rules for access to datasets and data sharing by individual repositories, we will introduce services and tools for the development of NDI in the processing of sensitive data. The purpose of this is to provide NDI users (researchers, research groups, etc.) with a secure environment in which they can process sensitive data. These sensitive data can be retrieved or stored directly from repositories in NDI. We focus on the development and implementation of services that:

1. enable the implementation of a secure process for receiving and releasing sensitive data to/from the environment for their secure processing (e.g. to/from storage areas and repositories in NDI);
2. enable the implementation of a suitable user interface for working with these data in a given secure environment (e.g. in the form of special VPNs, virtual desktops (VDs) for interactive work, an interface for running batch workflows (WF), etc.).

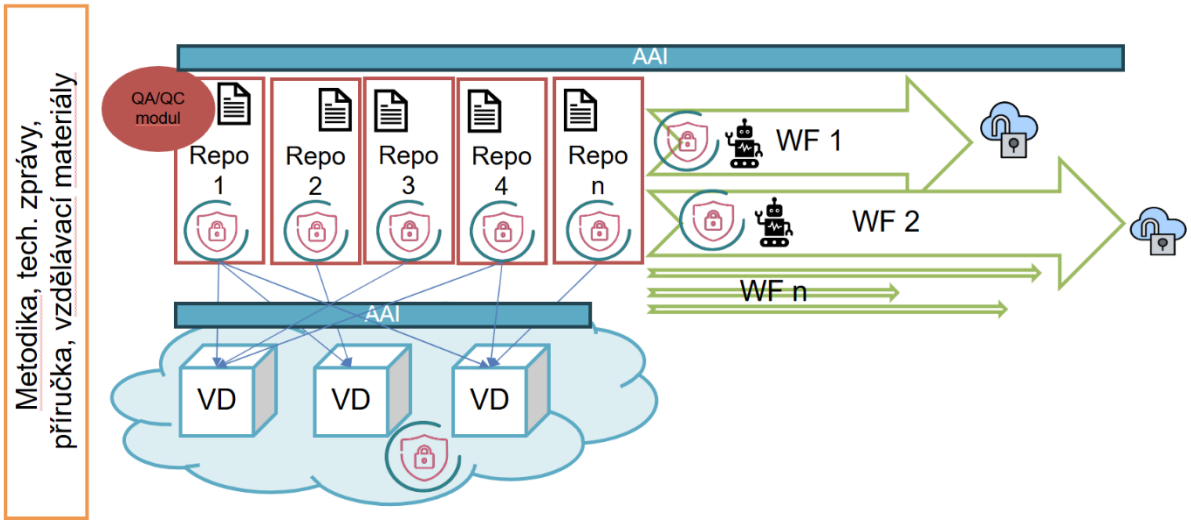

Fig. 12: Graphical diagram of TKA SENSI activities and outputs

This key activity will provide administrators and curators of NRP repository data with a software module for checking and ensuring the quality of repository data (QA/QC module[155]). TKA SENSI has decided to develop this module so as to be as universal as possible; however, the module will be developed on the basis of experience in the field of QA/QC clinical studies and biomedical research. Administrators, curators and users will have access to this SW module via REST-API[156].

10.8.1. SUB-ACTIVITY 9.1 – CREATION OF METHODOLOGY AND TRAINING MATERIALS FOR THE FAIRIFICATION OF SENSITIVE DATA157

The aim of this activity is to ensure that all repositories containing sensitive data meet the highest data protection standards and are fully interoperable with NDI. We will create a general methodology covering the standards and minimum requirements for the quality of (sensitive data) in the NRP, ethical and legal management and processing of sensitive data, covering also intellectual property and cooperation with the private sphere. As part of this methodology, links to the tools and services created by this TKA will be provided below.

Management and storage of sensitive data in the NRP, or NDI is currently not addressed at all from an ethical and legal, as well as data quality, point of view. We expect that the current situation, which is focused on a purely technical solution for the management of potentially sensitive data, will change thanks to the implementation of the Methodology to one that is clear and comprehensible for all entities that will want to store, request and process data of varying degrees of sensitivity in repositories. This will include cooperation with both academic and private entities.

We will prepare a general methodology for the management of sensitive data in repositories in accordance with ethical principles/standards for research and open science and with applicable European and national legislation, including the establishment of rules and definition of conditions for accessing and releasing data for further use in science and research.

The key topics in the FAIRification of sensitive data include: a secure environment for processing sensitive data; recommendations for security measures such as the encryption of data during transmission and in the repository; regular audits; establishment of rules for accessing data; and the management of sensitive data in a secure environment. This will include a description of the settings of sufficient authentication mechanisms and technologies for the detection and prevention of unauthorized access, in connection with the outputs of the NRP project, as well as a description of the workflow settings for the management of sensitive data with regard to their subsequent use. The issues of hiding sensitive metadata, storing relevant data in the provenance of information, and at the same time how to let users of the repository know about the existence of sensitive metadata will be elaborated here. We will focus on the design of a general technical solution for a repository that is suitable for the secure storage of sensitive data and, last but not least, proposals for ethical and legal means for sharing sensitive data, including the selection of a licence for datasets, linking datasets across repositories, data quality, etc. The general methodology will be shared for consultation within the project and applied to NRP repositories containing sensitive data.

[155] QA = quality assurance; QC = quality control.
[156] REST = REpresentational State Transfer; API = Application Programming Interface; https://restfulapi.net/.
[157] In the description of the KA, the planned activities are linked to the implemented outputs. The outputs are marked with a code. For TKA SENSI, this is in the format **SE_x** (key activity), **SE_x.x** (partial, i.e. part of the key activity).

The part of the methodology focused on ethical and legal aspects of data management will include:

- sample documents for data transfer agreement (DTA) and accessibility usage policy (AUP), sample queries to applicants for sensitive data regarding the planned use of the required data;
- ethical, legal and formal requirements for the long-term storage of research data and metadata;
- ethical, legal and formal requirements of applications for the re-use of data and their release for re-use;
- access to research data depending on the type of analysis performed (possibility of using anonymised, pseudonymised, de-identified and identifiable data) from an ethical and legal perspective;
- analysis of the involvement of the Data Access Committee (DAC) in the process of releasing data for re-use, including its composition, decision-making criteria, rights and obligations;
- design of an access pipeline (workflow) based on the above analyses;
- ethical and legal analysis of the use of a single identifier for individuals within repositories managing sensitive (biomedical, technical, technological, sociological and other) data, i.e. the possibility of linking personal data at the individual level between repositories;
- ethical evaluation – benefit versus risk analysis with respect to fundamental human rights, in particular the right to privacy.

The final document will be available online with the possibility of (non-)scheduled revisions and updates in the post-project period.

The methodology will be linked to specific repositories in the field clusters Bio/Health/Food, Social Sciences, Environmental Sciences, Data Management 4AI, Material Sciences, and repositories created in the project. Specifically, these repositories will process sensitive data. All repositories that work with sensitive data will be contacted as necessary by a bioethicist, an expert on cooperation with the private sphere, a lawyer, an AAI expert, and a data steward, all of whom will be coordinated and assisted by the TKA guarantor[158].

- National Omics Repository – Czech Omics Node (OmiCZ)
- repository for human and animal image and physiological multimodal data;
- a new CSDA repository for sensitive data for social sciences;
- ClinData repository;
- repository for storing data from non-targeted mass spectroscopic analyses for human exposome assessment;
- DM4AI repository;
- DANTE^{c} repository.

The analysis of requirements and preparation of the draft Methodology will be the responsibility of the data analyst (experts with a specific focus) and the methodologist (who knows partners, and processes, and understands systems and data), bioethicists, lawyers and TKA guarantors (specific

[158] The TKA SENSI team includes expertise on EHDS (the European Health Data Space Regulation valid from March 2025, further information available at https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_cs) or the preparation of implementation documents for this new regulation, as well as GDI (the Genomic Data Infrastructure Project, which seeks to create a federated, sustainable and secure infrastructure for access to genomic and related phenotypic and clinical data across Europe). Furthermore, members of TKA SENSI are involved in the activities of VVI BBMRI-ERIC, EATRIS, EBRAINS, Bioimaging, ELIXIR, EOSC and others.

with regard to the type of research data, type of cooperation and specialization) based on expertise in the FAIRification of sensitive data from the life stage of data, from preparation for access in repositories to sharing under clearly defined conditions. Supporting materials and documentation will be developed to create secure environments for the administration of sensitive data. While an analysis of good practice will be carried out as part of the preparation of the methodology, it will constitute a separate output and will encompass multiple levels of good practice for the sharing of research data obtained from collaborative research with the private sector and research of domain clusters involved in the NRP.

We plan the following levels of analysis:

1. Needs of partners in research and specific aspects by domain in order to ensure that the methodology for the FAIRification of sensitive data is as general as possible. We will take into account specific aspects of specific repositories and collaborations, e.g. analyses for the needs of sharing technical data from collaborative research will be carried out in a different way than for sociological or clinical data. With technical data, it is also assumed that different types of data will be handled differently, e.g. space research will be handled differently to, for example, energy.

2. Sensitivity of metadata in repositories – in some cases, not only the data will be sensitive, but also the metadata (specific cooperation, the content of which may be, for example, security research).

3. Types of sensitive data, including techniques for creating and handling synthetic data. At this level, it is also necessary to consider methods of anonymizing sensitive data, on which we will cooperate with TKA SOC, as well as the associated risks of reidentification. Alternatively, it is possible to build on the pilot "mini-project" for anonymization, which will deal with research into available methods, recommendation of procedures that sufficiently anonymize sensitive data, and the practical testing of healthcare data.

4. Need to establish a methodology in the process outside the domain repository with regard to some intellectual property issues.

5. Ethical and legal aspects of the management of sensitive data of human origin (e.g. sociological research, biomedicine – genomics, speech, image, clinical data, etc., use of artificial intelligence) **SE_1.**

As part of the preparation of the methodology for the FAIRification of sensitive data for the NDI, we will work with the results of the ongoing analysis and will draw on the experience of an expert team. At this stage of our work, general procedures arising in the process of creating a domain repository will be established, e.g.:

- determination of workflow for data and metadata retention (conditions – formal requirements, level of needs according to sensitivity, retention period, anonymisation approaches, handling of metadata, protection of trade secrets, etc.);
- determination of workflow for data access (from open access to the process involving specific access control for the handling of classified information, environment for processing, legal aspects and contractual documentation);
- determination of variants of licensing conditions when depositing data;
- determination of the roles and rights of user groups in the repository in individual parts of the workflow.

We will work on licensing conditions governing access to sensitive data in the NRP, including for automated virtual systems (bots), such as web crawlers[159], artificial intelligence (AI) machine learning and others; currently, there has been no in-depth analysis of these issues and it is not clear how they should be dealt with. We will focus on the preparation of documents for the management of sensitive data generated in cooperation with the commercial sphere and applied research. At this stage of the preparation of the Methodology, we will work closely with the teamworking on Compliance in the cross-sectional key activity of the OS II project and with the team for the relevant activity in the NRP project; we will utilize their interim reports and documents.

We will incorporate individual recommendations and expertise from specific areas and repositories regarding the processing of sensitive data into the methodology. Together with the administrators of repositories containing sensitive data, we will define standards and processes for managing sensitive data in their field. We will explicitly deal with the repositories listed above. To a greater extent, especially in connection with the operation of the CSDA repository[160] and the anonymisation of sensitive data from sociology or based on the examples of expertise in the automatic generation of metadata and data harmonisation in the ClinData system[161], cooperation and participation with other repositories from TKA B/H/F – Multiomic Human Data (OmiCZ) and the Repository for Human and Animal Image Data. In all cases, we will utilize the expertise of our team in the field of ethical issues of biomedical research within large research infrastructures and ethical issues of genetic testing and genomic data. Our TKA's expertise in cooperation with the private sector will be utilized primarily with regard to applied research, with a focus on intellectual property. In cooperation with the repository of material sciences, we will establish processes that will assist in the creation of technologies and related data management in practice.

The result of the cooperation described above will be the first version of the Methodology, which we will further iteratively modify in cooperation with project partners and test on the provided datasets.

As part of the preparation of the pilot version of the methodology, the established procedures will be tested on specific datasets of domain clusters within the TKA miniconsortium:

- Material Sciences – technical data; the goal of testing will be to verify the methodology regarding collaborative research;
- Bio/Health/Food – large image data and sequencing data of human subjects; testing will focus on the ethical and legal section of the methodology, because it is a special category of sensitive data;

[159] According to https://www.elastic.co/what-is/web-crawler, a data crawler is defined as "a digital search engine bot that uses metadata to discover and index web pages. Also referred to as a spider bot, it 'crawls' the world wide web (hence 'spider' and 'crawler') to learn what a given page is about. It then indexes the pages and stores the information for future searches."
[160] https://archiv.soc.cas.cz/cz/.
[161] ClinData is a complex information system developed by IMTM. It is designed to manage data in clinical trials, registries and other clinical or scientific databases. The system is characterized by a client/server web architecture that allows users to access it via web browsers on computers, tablets or smartphones. All communication between the client and the server is secured using SSL encryption, which ensures data security. ClinData enables the collection of clinical and laboratory data in accordance with GDPR from patients involved in various research projects and clinical trials. It supports the design of studies, the creation of electronic case report forms (eCRFs) and the management of study events or visits. The system also allows the storage of large scientific datasets, including genomic and proteomic data, by connecting to the IMTM object storage infrastructure. In addition, ClinData facilitates the FAIRification (findable, accessible, interoperable, reusable) of data by exporting selected datasets to public repositories. This system will be implemented in the NRP as part of TKA B/H/F as a repository for clinical data.

- Social Sciences – the goal of testing will be to verify data anonymization techniques when sharing, which need to be established specifically according to the required data and the purpose for which they are requested.

The resulting methodology will cover the recommended points (known as the checklist) that a sensitive dataset should meet before being accepted into the repository in order to meet the requirements for sensitive data and their sharing. We will also prepare instructions for applicants for datasets containing sensitive data, how to request data and how to handle it in a secure environment for the processing of sensitive data or after receiving it from the repository. At the end of the project, recommendations and standards for the handling of sensitive data in the NDI will be clearly defined in the form of a Methodology covering the ethical and legal handling of sensitive data and their quality during their FAIRification. This will include best practices from the preparation of data for uploading to a sensitive data repository, through retention in the repository, to the conditions and method of sharing sensitive data of different categories. These outputs will be available not only for the needs of NRP users, but also for communities that operate repositories outside the NRP, in order to ensure the necessary compliance in the event of subsequent need. This activity will include consultations and expert advice for emerging repositories that store sensitive data. We will work closely with the NRP compliance team and the cross-sectional activities of OS II Compliance, Cybersecurity, Provenance and Language Models. Currently, there is no comprehensive methodology dealing with this topic. **SE_2**

Training materials based on the Methodology for the FAIRification of Sensitive Data in the NDI will be created in order to achieve a more systematic dissemination of know-how. The material will provide the necessary knowledge for the FAIRification and management of sensitive data in the NDI as a whole, including sample documents, links to domain repositories, National Metadata Directory (NMD), legal frameworks for the FAIRification of sensitive data, ethical aspects and more.

The content of the methodology will be presented to NDI users and all those who would like to create a new repository in the NRP via dedicated NDI/NRP pages.[162] **SE_2.1**

The activity aimed at creating a methodology for the FAIRification of sensitive data will include trips abroad for the purpose of gaining knowledge of and experience with the management of sensitive (research) data from other countries, research infrastructures and communities. Dedicated team members, especially those who will create the methodologies or services/tools for NDI, will be sent for about two trips of several days in length (seminars, conferences, congresses, etc.).

ACTIVITIES:

- analysis of good practice both in the Czech Republic and abroad for sharing research data obtained from collaborative research with the private sector and research of domain clusters involved in the NRP, including ethical and legal analysis; **SE_1**
- preparation of a draft comprehensive Methodology for the FAIRification of Sensitive Data within the NDI by experts of the implementation team in cooperation with administrators of repositories containing sensitive data;
- internal verification/testing of the procedures described in the Methodology (passage of the draft workflow) on selected datasets;
- finalization of the Methodology based on interactions with repositories; **SE_2**
- creation of training materials based on the Methodology for the FAIRification of Sensitive Data. **SE_2.1**.

[162] The training courses themselves will be realized in cooperation with the EOSC CZ Training Centre, which will provide adequate capacities, technical facilities and organizational support and are not included in the budget for OS II.

SUB-ACTIVITY OUTPUT CODES[163]

SE_1	Analysis of good practice for the sharing of sensitive data from collaborative research
SE_2	Methodology FAIRification of Sensitive Data in NDI
SE_2.1	Training materials based on the Methodology for the FAIRification of Sensitive Data

10.8.2. SUB-ACTIVITY 9.2 – DEVELOPMENT AND PILOT IMPLEMENTATION OF SERVICES AND TOOLS FOR THE DEVELOPMENT OF THE NDI

In this activity, we will focus on the development and implementation of tools and services that will enable the implementation of a secure process for receiving and issuing sensitive data to/from the environment for their secure processing, e.g. to/from repositories and repositories in the NDI; as well as the implementation of a suitable user interface for working with these data in a given secure environment, e.g. in the form of dedicated VPNs, virtual desktops (VD) for interactive work, an interface for running batch workflows (WF), etc. The purpose of this is to provide NDI users (researchers, research groups, etc.) with a secure environment in which they can process sensitive data.

As part of the activity, we will draw on expertise on the European Health Data Space Regulation (EHDS)[164], or the preparation of implementation documents of this new regulation, as well as experience and documents from GDI, etc.[165] The solution will incorporate the experience and know-how of specialists on the processing of sensitive data in the SensitiveCloud CERIT-SC solution[166]. The experience of part of the team with the administration of sensitive (clinical) data on the ClinData platform will be essential for the development of the (clinical) data quality control tool. An API will also be created to allow external clients to access the Clindat QA/QC module.

The goal of the activity is to ensure the control and quality assurance of data in individual repositories based primarily on the ClinData system so that the dataset meets standards such as ICD 10[167][168], LOINC[169], OMOP, ISO 8601 for the date[170], or others, as required. The current form of the[171] ClinData system is already being used in a number of hospital facilities for multicentre clinical trials and can be considered a standard tool for the collection of parametric clinical records. Modifications to the ClinData system to allow it to serve as a NRP repository are dealt with in TKA B/H/F.

This activity also includes trips abroad, which are necessary for the acquisition of experience with tools and services for working with sensitive research data at national and international level. The purpose of the trips will be in particular to attend thematic conferences, seminars, workshops, or short

[163] The sub-activity output codes are linked to Annex 3 Schedule of Key Activities and chap. 11 Main planned outputs/products of Feasibility Study.
[164] The European Health Data Space Regulation is a new EU regulation that has been in force since March 2025. Further information is available at: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en.
[165] Team members are also involved in the activities of VVI BBMRI-ERIC, EATRIS, ECRIN, EBRAINS, Bioimaging, ELIXIR, EOSC, and others.
[166] Cf. https://www.cerit-sc.cz/infrastructure-services/sensitivecloud.
[167] https://icd.who.int/browse10/2019/en.
[168] https://loinc.org/.
[169] https://ohdsi.github.io/CommonDataModel/.
[170] ISO 8601-1:2019/Amd 1:2022 - Date and time — Representations for information interchange — Part 1: Basic rules — Amendment 1: Technical corrections.
[171] ClinData is a comprehensive information system designed to manage data in clinical trials, registries and other clinical or scientific databases.

internships in institutions engaged in the development/operation of tools and services for working with sensitive research data. The outputs of this sub-activity will also be actively presented at least at one international event.

Software module for checking and ensuring the quality of repository data

No system is currently in place for the automatic examination of the quality of data when uploading to NRP repositories and the only verification process in place is verification by the repository curator; this is unsustainable in the long term for reasons of capacity and human error. In order to meet the needs of data administrators and curators of NRP repositories, we will develop a QA/QC software module for checking the quality of uploaded data in order to eliminate data errors and ensure the checking of the relevance of data within the repository. Due to the fact that the expertise of the QA specialist and programmers in the team involved in the development of the module lies in the administration of clinical data, the QA/QC module will primarily be created for the ClinData system. In this way, we will utilize expertise from clinical medicine for other fields. ClinData-based repositories will use the QA/QC module automatically. Access to this QA/QC module for repository users will be implemented via an API. This interface will allow administrators of other types of repositories to implement the QA/QC module into their environment.

Basic QA/QC module functions

The proposed QA/QC module will be a software system enabling the validation of an uploaded dataset according to defined rules. While validation rules will be based on existing harmonization standards (OMOP, etc.) or ontologies, it will be possible to define special rules on the basis of expert assessment. The output of the QA/QC module will be the subsequent report, with a list of errors and warnings in the analysed dataset. The QA/QC module will be usable in the ClinData system, but also in other repositories thanks to the implementation of the API programme interface. The results of the validation will be shared with the curator of the repository into which the module will be implemented. The specific use of the functionality and establishment of validation rules of the module will be up to the administrator or curator of the repository data according to the domain.

While developing the module, we will follow standard agile SW development procedures, i.e. we will perform a needs analysis, divide development of the tool into individual parts (user stories) and create basic software components and programme interfaces. We will design test scenarios to verify the functionality of the module, and will also test the user-friendliness of the graphical interface. We will implement the software iteratively. The basic architecture of the QA/QC module will consist of the following components:

- **Input module** in the form of an API or web interface. This will allow the uploading of the dataset into the QA/QC module environment and access to individual items of inserted data. The data will be imported in text form; we anticipate that the csv format or similar will be used.
- **The data analyser** will allow the dataset to be loaded into the data structure of the QA/QC module and connect individual parameters with the harmonization standard or ontology.
- The **designer of validation rules** will be operated via a web interface, where the user defines validation rules for individual parameters and links between parameters. Reusable validation rules can only be created for harmonized parameters.
- The persistence of validation rules will be ensured by using the SQL/NoSQL database.
- The **validation procedure** will allow the running of validation rules over the inserted data. The report will be created from the resulting set of findings.

- The **reporting module** creates an output report from the validation procedure. The report will be obtainable from the web interface or the API programme interface.
- The **authorization module** will allow controlled access to the web interface. Access will be controlled at the level of datasets and validation rules. It will be linked to the NRP AAI[172].
- **Security** will be an integral part of the API programme interface; we anticipate the use of client certificates to connect to the API.
- **Logging and auditing section**.

The above validation procedure will work **with validation rules that will work with both data and metadata at multiple levels:**

- Basic validation of an individual parameter – filling check of meta/data, format, scope, etc.
- Advanced validation of an individual parameter according to the harmonization standard – checking whether the entered values correspond to the definition, reference range, etc.
- Advanced validation of a group of parameters – checking of links between dependent parameters; these are checks of the type "correct sequence of events", "parameter value excludes a state", "conditional obligation of a parameter", etc.
- Plausibility – checking the credibility of the data will depend on the expert assessment and the weighting of individual validation rules.

Validation rules will be configurable according to the needs of the repository using the validation rules designer. The output of the validation procedure will be a set of findings of differing severity; we anticipate the following basic levels: OK/information/warning/error. Testing and the pilot run of the module in real operation will be carried out in cooperation with TKA B/H/F and the clinical data repository.

The QA/QC module development process will also include the creation of a production environment for deployment. We will carry out random user (exploratory) testing, correct errors and ensure the operation of **the module in the production environment.**

The security of the QA/QC module will be dealt with on several levels.

- Security of access to the web interface. Every user will have to log in via a central authentication service connected to the NRP AAI. The connection to the web interface will be secured by SSL encryption and a certificate.

- The security of the connection to the API programme interface will be implemented using a client certificate. However, this may still be reassessed.

- Securing the dataset. The QA/QC module will not be used for the long-term storage of validated data. We anticipate that, after performing the validation, the data will be released from the module's temporary storage area after a defined period of time. Every user will see only their own datasets in the module interface; we do not anticipate the sharing of data between users. For this reason, it will not be necessary to consider the security of datasets at the level of individual records. It will be possible to insert a dataset with sensitive data into the QA/QC module; security will be at a level appropriate for sensitive data.

- The validation rules created will be saved persistently and it will be possible to share them between users.

[172] NRP AAI = NRP authentication and authorization infrastructure; utilizes a user management system called Perun and is used to control user access to individual NRP resources.

The technical documentation for the QA/QC module will also include a methodological manual for data curators and a practical manual with training material. A detailed description of the API programme interface for data input and output from the QA/QC module will also be created, enabling the implementation of this module in other repositories intended for processing parametric data. We anticipate use of the following **technologies and software architecture for the QA/QC module**:

- We expect that the QA/QC module will be written in the Java programming language (version 21 or later).
- We expect that Spring Boot will be used as the basic software framework for the backend.
- During the development process we want to emphasize the modularity of the source code based on programme interfaces. It is anticipated that the primary data input and output will be the REST API. We expect to verify the functionality of all modules with a set of unit tests, which we expect to write in parallel with the source code. The goal is to cover as much of the source code as possible with tests.
- We expect that the test framework will be JUnit.
- We anticipate that jdbc, Hibernate, etc. will be used to ensure data persistence (configuration, validation rules, etc.).
- We anticipate using the SQL database for data storage.
- The source code will be accompanied by detailed comments at module and function level.
- We assume that the client component of the QA/QC module will be written in Java script and the use of a framework such as jQuery, Angular or React.
- We anticipate that the source code will be saved in the GIT distributed versioning tool and the entire development process will be governed by the principles of Continuous Integration (CI), i.e. development will be divided into small units (user stories/tasks in the Jira system), which will be released gradually. The disintegration of the entire QA/QC module into these small pieces must be preceded by a more detailed analysis and is not included in this text.
- A development and testing environment will be created in which it will be possible to test the developed modules/parts of the code and their functions. **SE_3**

ACTIVITIES:

- analysis of the development of the QA/QC module and the breakdown of the whole into individual parts (user stories/tasks);
- development of a software QA/QC module by individual components;
- pilot operation of module on the clinical data repository (TKA B/H/F, ClinData repository);
- automated unit tests, check and validation of the source code;
- creation of an API programme interface to allow access by external clients;
- creation of a production environment, deployment, random user (exploratory) testing, error correction, security of the production environment, optimization of operation;
- creation of related module documentation. **SE_3**

Services for the development of the NDI

The goal of this activity is to effectively connect existing environments for the processing, storage, and acquisition of sensitive data (such as SensitiveCloud or ClinData) to/from the NDI. This encompasses possibilities for the technical implementation of a controlled process for the receipt/release of data to/from such a sensitive data processing environment, the implementation of appropriate strong data protection during transmission, and innovative end-user interfaces for interactive/batch

processing of sensitive data in a given secure environment. We will introduce four services. As part of this activity, existing tools and environments will be developed into more advanced forms (e.g. virtual desktops), or standardized forms integrating existing solutions (e.g. REMS).

1/ Service for a controlled, secure process for the receipt/release of data to/from the sensitive data processing environment

This service is not linked to a specific repository; the process can control the receipt/release of data from/to any source, including NDI repositories. It also supports activities necessary for the FAIRification of data in repositories, or the processing of sensitive data. It also supports the implementation of Trusted Research Environment system services.

This service will implement a secure controlled process for the receipt/release of data to/from a sensitive data processing environment. It will therefore, for example, enable the authentication of data providers/recipients, allow the definition of the data process before receipt/release, e.g. the verification of quality before receipt, sufficient aggregation/anonymization before release, etc. This will enable the implementation of a secure interface between secure and potentially risky environments for sensitive data.

Implementation will be based on an analysis of existing recommendations, best practices, standards, and requirements or existing tools for the receipt/release of data to/from a sensitive data processing environment.

The output will be a standard component of, and will continue to be maintained on, one of the sensitive data processing environments. At this moment in time, we anticipate implementation in the SensitiveCloud environment[173]. We expect the NDI infrastructure (e.g. repositories) to provide and maintain standardized interfaces that will be usable for interconnections.

Documents will be collected on an ongoing basis and a technical report will be prepared with recommendations for a controlled secure process for the receipt/release of data to/from a sensitive data processing environment. Together, they will constitute technical and methodological documentation that will enable the implementation of similar services in other environments for processing sensitive data in the context of the NDI.

ACTIVITIES:

- analysis of existing requirements, recommendations, tools for the receipt/release of data to/from a secure environment;
- implementation of a service for the receipt/release of data into/from a secure environment;
- further deployment and operation of the data receipt/release service in the SensitiveCloud environment;
- collection of documents and preparation of a technical report entitled "Recommendations for a controlled secure process for the receipt/release of data to/from a sensitive data processing environment";
- finalization of the service and its documentation link to the output. **SE_4**

[173] Cf. https://www.cerit-sc.cz/infrastructure-services/sensitivecloud.

2/ End-user interface for the interactive/batch processing of sensitive data in a secure environment

This service allows the processing of sensitive data, e.g. processes to monitor/improve their quality. It also supports activities necessary for the FAIRification of data in repositories, or the processing of sensitive data. It also supports the implementation of Trusted Research Environment system services.

As part of the implementation of this service, appropriate interfaces for the interactive processing of sensitive data in a secure environment will be identified according to an analysis of user needs, and these will be implemented. For interactive processing, we are considering, for example, implementation in the form of isolated virtual desktops or line terminals, and for batch data processing we are considering the use of a suitable tool for process orchestration (e.g. Snakemake, etc.). At this moment in time, we anticipate deployment in the SensitiveCloud environment[174].

Documents will be collected on an ongoing basis and a technical report prepared, which together will constitute the technical and methodological documentation that will enable the implementation of similar services in other environments for processing sensitive data in the context of the NDI. The technical report together with the technical solution for the service described above will enable the implementation of similar services in other sensitive data processing environments.

ACTIVITIES:

- analysis of user needs for the interface for processing sensitive data;
- implementation of a user interface service for the interactive processing of sensitive data;
- implementation of a user interface service for the batch processing of sensitive data;
- deployment and operation of a user interface service for the interactive processing of sensitive data;
- deployment and operation of a user interface service for the batch processing of sensitive data;
- collection of documents and preparation of a technical report entitled "Recommendation on interfaces for end users for the interactive/batch processing of sensitive data in a secure environment";
- finalization and documentation of the service **SE_5**

3/ Service for the authorization of the batch processing of sensitive data

The output of the service is the establishment and implementation of a mechanism that will allow the authorization of requirements for the batch processing of sensitive data in a Trusted Research Environment. Every request will contain information about the analysis performed and its authorization through the AAI infrastructure. This service will allow the performance of analyses on sensitive data without the need for direct user access to the data itself, which is a necessary first step in solving the problem of federated calculations and the consortium sharing of sensitive data analyses. The primary goal of the service is the Omics Repository (OmiCZ), which contains sensitive genomic data, but is also compatible with and applicable to any other sensitive data repository that needs to run analyses without direct user access to the data itself. A technical report, which will constitute the technical and methodological documentation of the output "Interface for end users for interactive/batch processing of sensitive data in a secure environment" described above, will also be prepared. This will enable the implementation of similar services in other sensitive data processing environments.

[174] Cf. https://www.cerit-sc.cz/infrastructure-services/sensitivecloud.

The output will be a standard component of, and will continue to be maintained on, one of the sensitive data processing environments.

ACTIVITIES:

- collection of workflow authorization requirements, definition of use scenarios. Mapping of type analyses, proposal of a method for transferring information about the analysis and its authorization via AAI;
- design of service architecture, design of a methodology for authorizing batch processing, integration with AAI, definition of request formats;
- implementation of a service prototype, testing on pilot scenarios in the OmiCZ environment and continuity with the environment for processing sensitive data;
- finalization of the service, preparation of technical and methodological documentation (technical report). Commissioning of the service **SE_6**

4/ Service for the evaluation of analytical workflows and classification of the sensitivity of their outputs

This service will provide a systematic evaluation of the sensitivity of the outputs of analytical workflows. The service will be implemented as a semi-automatic system that will combine predefined rules, tools for monitoring the structure of outputs and expert evaluation (in cooperation with experts involved in the development of the OmiCZ repository – specifically the research data curator, the OmiCZ repository administrator and the DevOps specialist for sensitive data analysis). The output will also include sample and template workflows that will be validated in terms of the sensitivity of the outputs and usable as reference solutions for consortia or research teams that need to share only certain types of data – e.g. aggregated or desensitized data. The service will thus provide a key basis for establishing access rules and secure data sharing both within and between research infrastructures. The primary goal of the service is the Omics Repository (OmiCZ), which contains sensitive genomic data, but is also compatible with and applicable to any other sensitive data repository that needs to run analyses without direct user access to the data itself.

A technical report will be prepared presenting the methodological and technical documentation for the "Service for the evaluation of analytical workflows and classification of sensitivity of their outputs". It focuses on the issue of determining the degree of sensitivity of output data from analytical processes, especially from the point of view of compliance with the GDPR and the possibility of their secure sharing. The report includes a description of a semi-automatic system that combines predefined rules and control of the structure and typology of outputs with expert input. The report also includes a set of sample and template workflows that have been validated for output sensitivity and are intended as reference solutions for consortia and teams that plan to securely share aggregated or anonymized data. The goal of the documentation is to facilitate the transferability of the methodology to other repositories or environments for processing sensitive data within the NDI and to support a uniform approach to the classification of output data sensitivity.

ACTIVITIES:

- draft methodology for assessing the sensitivity of outputs, analysis of community requirements, definition of sensitivity rules and type scenarios;
- design of architecture for a semi-automatic system for sensitivity assessment, design of sample and template workflows;

- implementation of a system for sensitivity assessment, validation and testing on real scenarios in the OmiCZ repository/environment for processing sensitive data;
- finalization of the service, preparation of technical and methodological documentation (technical report), ensuring of portability to other NDI environments. **SE_7**

SUB-ACTIVITY OUTPUT CODES[175]

SE_3	SW module for monitoring and ensuring the data quality of repositories incl. methodological manual of QA/QC data in repositories
SE_4	Service for a controlled, secure process for the receipt/release of data to/from the sensitive data processing environment
SE_5	End-user interface for the interactive/batch processing of sensitive data in a secure environment
SE_6	Authorization service for the batch processing of sensitive data
SE_7	Service for the evaluation of analytical workflows and classification of the sensitivity of their outputs

10.9. CROSS-SECTIONAL PROJECT THEMES

10.9.4. SUB-ACTIVITY 10.4 – ELECTRONIC LABORATORY NOTEBOOKS FOR THE NRP[176]

Electronic laboratory notebooks (ELNs) are one of several types of software tools used by researchers to manage and share scientific data, and prepare it for publication. This category also includes tools referred to as electronic field notebooks[177], which focus on domains, including field work (e.g. archaeology, ecology, field biology).

A high-quality ELN provides researchers with a user-friendly environment in which they can pair data obtained, e.g. from an experiment or calculation, with a description of them (metadata), safely store them, search for them, control their sharing with other researchers, and monitor their life cycle. In many cases, ELNs allow the collection of data from complex measurements or complex research processes. This can result in data that are fully annotated and linked.

ELNs focus on working *with hot and warm data*[178]. They aim to ensure the quality of data (including the implementation of FAIR principles) from the moment of their creation, through recording the process of their processing and enrichment with additional data or metadata, to versioning in the case of repeated or supplemented outputs.

The main function of ELNs is advanced work with metadata. Descriptions of data (metadata) can be developed into any level of detail specific to a given scientific discipline or a given project. At the same time, this process is fundamentally simplified in the ELN using forms and pre-filled data in order to

[175] The sub-activity output codes are linked to Annex 3 Schedule of Key Activities and chap. 11 Main planned outputs/products of Feasibility Study.
[176] In the description of the KA, the planned activities are linked to the implemented outputs. The outputs are marked with a code in the format **ELN_x**.
[177] This is an identical type of SW tool as the Electronic Laboratory Notebooks,
[178] not like data repositories within the meaning of the NRP project. The NRP project does not offer tools for the advanced management of *live data* that would be directly deployable in scientific organizations. The NRP project focuses on building infrastructure. ELNs are mainly used in technical and scientific disciplines, and some aspects of ELNs (e.g. metadata forms) are domain-specific.

record information that repeats itself. ELNs make it possible to unify data processing procedures so that data and metadata are consistent across different experiments and different personnel collaborating on research, including harmonization across institutions or international partners. In some cases, ELNs allow the visualization of stored data or the addition of schematic records, including complex chemical or mathematical equations, using integrated tools directly in the ELN. From the point of view of legislation, some ELNs allow the legally valid identification of the origin of data (signatures, witnessing, timestamping).

Another important native feature of the ELN is the recording of the history and development of and changes to data (provenance). This naturally builds on sub-activity 10.5, which deals with provenance. The synergy of this link lies in the fact that sub-activity 10.5, with its methodology, will provide general guidance on how to properly use ELNs so that users respect the legal framework of provenance. The software tools developed in PKA Provenance also offer an alternative to provenance management in situations where ELNs will not be used – they are complementary. For NRP/NDI purposes, ELNs are a recommended tool for the consistent, user-friendly collection and description of data, including their export to data repositories.

- Internationally, ELNs are used in scientific fields corresponding to several TKA OS II projects (e.g. HUMAN, PHYSICS, MATECH, ENVIRO or BIO/HEALTH/FOOD). They are considered a standard tool for managing hot and warm scientific data and preparing them for publication in trusted FAIR data repositories. Due to the compatibility of Czech research groups with the international environment and their competitiveness, ELNs must be included in the portfolio of NDI software tools supported at the national level.

- In the domains of MATECH, PHYSICS and Archaeology (part of HUMANITIES), ELNs have been identified as critical tools for creating FAIR data and preparing them for publication in trusted NRP/NDI repositories. It can be assumed that for some of the sub-clusters of the BIO/HEALTH/FOOD and ENVIRO clusters, these tools will also be introduced in the near future for the management of research data due to the aforementioned competitiveness at international level and the need to process complex datasets (see below).

- The quantity of research data and related metadata is also growing rapidly across domains. Evolving data processing technologies, such as machine learning, then allow increasingly complex data workflows. This places increasing demands on data management, for example due to the registration of detailed metadata, provenance, or standardization across scientific disciplines. ELNs are a type of tool that can meet and simplify these increasing demands.

The deployment of ELNs on a national scale in the NDI environment requires that some of the objectives of the NRP project (AAI, data repositories, data repositories, etc.) have already been met. The implementation of this area is closely related to the existence of domain-specific repositories and it is necessary that the scheduling of the area be tied to these. The following interconnections are therefore in place within the project:

- In order to properly integrate the ELN with the NDI catch-all data repository, it is necessary for the research team to have a metadata model and related publication workflows. These are the outputs of the OS I project. Therefore, the completion of the ELN integration output with the repositories is scheduled only after the planned completion of the metadata model in OS I.

- Similarly, for the integration of the ELN with the NDI catch-all repository,[179] it is necessary for the production instance of the repository itself to be in operation. This specifies the SW system for which it will be necessary to create integration tools in the ELN and at the entrance to the repository. Therefore, the completion of the integration of the ELN with the catch-all NDI repository is scheduled only after the planned deployment of the production instance of the repository.

Based on a thorough survey of the different ELNs currently available, two were selected to be supported as part of the sub-activity: Kadi4Mat and eLabFTW. These two notebooks best meet several important criteria essential for long-term deployment in NDI:

- They were developed by a stable team of experienced developers. The probability of interruption of development in the project sustainability time horizon is very low.
- The development of these products is financed by large research organizations (KIT Germany, CNRS France), where they also have a broad user base.
- Both electronic laboratory notebooks are being developed as open-source software and have the potential to support FAIR data in the future.
- Both electronic notebooks selected have a simple intuitive user interface, small administration overhead and no specific features that tie them too closely to any one field of science.
- The selected ELNs are based on standard software technologies for which there is a high probability of support in the project sustainability time horizon. Both notebooks have a good REST API structure that allows their integration with other systems within the research organizations where they will be deployed, as well as in the NDI.

As part of the sub-activity, SW services will be provided for the integration of supported ELNs with NRP repositories within the NDI. The goal is for researchers to have a simple and user-friendly means of publishing ELN records to data repositories directly within the ELN environment without having to re-enter information or manually export/import data and metadata.

As part of the sub-activity, national instances of both supported ELNs (Kadi4Mat and eLabFTW), which will be intended for organizations that will not have their own ELN instances, will be deployed. The partner will then purchase, from the direct funds of the PKA ELN, a server on which these instances will be operated, and provide appropriate software support.

As part of the sub-activity, a methodology for the recommended deployment and operation of supported ELNs in organizations that will operate their own instances will be developed. It defines the minimum requirements for hardware, software, infrastructure and human resources, prepares documentation of the recommended deployment procedure for the supported ELNs in relation to the NDI/NRP.

As part of the sub-activity, new properties will be developed for the Kadi4Mat ELN, which will further expand its long-term application in the NDI. The main emphasis will be placed on the implementation of communication between different instances of the Kadi4Mat ELN, which will enable the interactive sharing of data between different organizations and research groups (including international collaborations). This feature of the Kadi4Mat ELN represents a frequent demand from future users of the ELN and is of great interest in the Czech research environment. Due to the fact that the Kadi4Mat ELN does

[179] Cf. repo.cz.

not currently directly support the deployment model in the container platform environment offered by the NDI, the partner will take steps to make it possible to deploy Kadi4Mat instances in the Kubernetes environment within the NDI at the end of the project.

As part of this sub-activity, the implementation team will support two specific ELNs developed by the scientific community: Kadi4Mat and eLabFTW.

The first mentioned notebook (Kadi4Mat) is very general and is suitable for a wide range of scientific directions. It can be effectively extended using its own modules, which can provide an advanced specific functionality that may be required by a specific scientific team. Kadi4Mat has excellent prerequisites for integration with scientific workflows and for use with systems for the automatic collection of scientific data from measuring devices[180].

The second supported notebook (eLabFTW) is more specific and its architecture is more consistent with recording the course of a scientific experiment over time. It is easier to deploy than Kadi4Mat and has specific functions relevant in sciences close to chemistry and molecular biology.

The joint support of both notebooks, which have complementary properties, aims to cover the widest possible group of users with diverse requirements.

As part of the sub-activity, a service will be implemented to allow users to publish data and metadata from records in the Kadi4Mat ELN to the catch-all NDI repository and to various domain data repositories in the NDI. The service will be implemented both in the user interface and on the backend.

The service will consist of two parts: one will be a Kadi4Mat component (plugin), and the other will be an independent microservice between the ELN and the repository.

The functionality of the output can be verified by accessing the national Kadi4Mat ELN instance, where it will be clearly visible in the user environment. At the same time, a public Git repository will be available with the relevant component source codes and development history.

We will implement the integration of the generic catch-all datarepo.eosc.cz repository in the NDI with the national instance of the Kadi4Mat ELN. Within individual TKAs, the service will be available and the integration of individual instances of the Kadi4Mat ELN with domain-specific NRP repositories in the NDI will be possible. **ELN_1**

In addition, a service that allows the user to publish data and metadata from records in the eLabFTW ELN to a generic catch-all repository and to various domain data repositories and NRPs in the NDI will be implemented. The service will be implemented both within the user interface and on the backend. The service will utilize a standardized API of the application and repositories. This will facilitate sustainability with minimal financial costs. The service will consist of two parts. One will be implemented as a component in the eLabFTW ELN. This is outsourced to the eLabFTW development team. The second part of the service is implemented as an independent microservice between the ELN and the repository. The functionality of the output can be verified by accessing the national eLabFTW ELN instance, where it will be clearly visible in the user environment. At the same time, a public Git repository will be available with the relevant component source codes and development history.

We plan to outsource the development of that part of the NRP integration service, which is a component of eLabFTW, to the Deltablot development team (France, NPO) – the authors of eLabFTW ELN. This is due to the software technologies used, on which the eLabFTW user interface is built. We have no experience with these technologies (PHP, specific architecture of the software product) in the team

[180] These are part of the KA5.3 project of the NRP.

working on the sub-activity. Therefore, it will be more time-efficient and economical to outsource the development of the component to the authors of eLabFTW. The authors of eLabFTW offer such a service, and at the same time are competent to handle the task at a lower cost as they have detailed knowledge of the internal architecture of the eLabFTW ELN. The second part of the service development (a separate microservice) will be developed in the PKA team.

We will implement the integration of the catch-all repository in the NDI with the national instance of the eLabFTW ELN. Within individual TKAs, the service will be available and the integration of individual instances of the eLabFTW ELN with domain-specific NRP repositories in the NDI will be possible. **ELN_2**

A methodology will be prepared for deploying the Kadi4Mat and eLabFTW ELNs in scientific organizations that will create and administer their own instance using NDI resources. Specifically, minimum hardware and software requirements will be specified based on the size of the organization and the quantity of data. In addition, instructions for the integration of ELNs with NDI tools (AAI, data repositories, container platform for providing virtual services) and instructions for scaling ELN deployments will be created. The output will also include instructions for organizing record templates so that they are compatible with the metadata models of the relevant NRP repositories in the NDI and a model template for recording metadata for the generic catch-all repository (CCMM). The output will be carried out from a position in the sub-activity. **ELN_3**

One instance of the Kadi4Mat ELN and one instance of the eLabFTW ELN will be deployed for organizations that do not have the technical capabilities or interest to set up their own instances of supported ELNs. Both parts of the NRP infrastructure in NDI (AAI, container platform, S3 data repositories) and the virtualization server purchased from direct funding for the sub-activity will be used. This is because, at minimum, the Kadi4Mat ELN is not yet compatible with the Kubernetes container platform[181] available within NDI. The deployment and administration of the instances will be financed from the personnel costs of the sub-activity. **ELN_4**

In cooperation with the development team at IAM KIT,[182] Karlsruhe, Germany, new properties of the Kadi4Mat ELN that are relevant in the national and European scientific environment will be implemented. Specifically, this will be the interconnection of different instances of the Kadi4Mat ELN. This will allow users, after authorization, to interactively share data between instances through reciprocal direct access, without export and import. This is important not only for large data, but also for records for which transfer between different ELN instances is impractical and such operations must be recorded in the records metadata (provenance). It will also allow different teams and different ELN instances to work together remotely when processing the same data. This approach is highly innovative in the world of ELNs because this complex feature is not typical of other ELNs.

A further element implemented in the project will be a notification system that will notify users interactively about events in the ELN and allow them to respond to them. Work on this output will also include customization of the application for easier routine deployment of the ELN instance in the Kubernetes container platform environment.[183] This will further facilitate the adoption of ELNs in other

[181] Most of the OS II project software components will run on NDI hardware resources. That provides access to the Kubernetes – Rancher container platform(https://www.cerit-sc.cz/infrastructure-services/data-processing/container-platform). Some software components planned for the activity are not compatible with this platform and will have to be operated outside the container platform. We therefore plan to purchase one server with hardware parameters suitable for running several virtual servers with these software components. Over the duration of the project, the one-time cost of purchasing a server and its operating costs is lower than the total cost of typical fees for renting virtual servers from external providers.

[182] Institute of Advanced Materials, Karlsruhe Institute of Technology.

[183] https://www.cerit-sc.cz/infrastructure-services/data-processing/container-platform.

research organizations on NDI funds.

This output assumes that the SW developer undergoes an annual short stay (five days) in the Kadi4Mat development team in Karlsruhe, Germany. The purpose of this is to synchronize activities and find out about the further conceptual direction of the system. Kadi consists of several components (Kadi4Mat ELN, Kadi Workflows, Kadi FS, Kadi AI, etc.), whose philosophy and concept of use is constantly evolving. For this reason, it is important that we keep in touch and monitor developments. It will also make it possible to communicate the needs of the target group[184] to developers and implement them more effectively. Therefore, the PKA will include a planned annual five-day business trip per year for one person to Karlsruhe, Germany for intensive in-person cooperation with the Kadi4Mat development team. For the rest of the year, the SW developer will work remotely with the Kadi4Mat development team. **ELN_5**

ACTIVITIES:

- integration of the Kadi4Mat ELN with the NRP; **ELN_1**
- integration of the eLabFTW ELN with the NRP; **ELN_2**
- deployment methodology for the Kadi4Mat and eLabFTW ELNs; **ELN_3**
- deployment of national ELN instances; **ELN_4**
- development of new properties for the Kadi4Mat ELN. **ELN_5**

SUB-ACTIVITY OUTPUT CODES[185]

ELN_1	Integration of Kadi4Mat ELN with the NRP
ELN_2	Integration of the eLabFTW ELN with the NRP
ELN_3	Deployment methodology for the Kadi4Mat and eLabFTW ELNs (WIKI chat-ready dock)
ELN_4	Deployment of "catch-all" ELN instances (1 instance of Kadi4Mat ELN, 1 instance of eLabFTW ELN)
ELN_5	Upgraded ELN Kadi4Mat

10.9.5. SUB-ACTIVITY 10.5 – TRACEABILITY OF OBJECTS USING PROVENANCE[186]

The outputs developed as part of this sub-activity will enable harmonized work with trusted provenance in[187] order to support the traceability of the predecessors of objects stored both within and outside repositories (or originating from laboratory notebooks or other SW). Provenance will be represented and administered in accordance with current international standards, in particular W3C[188] PROV and CPM[189]/ISO[190] 23494, in order to ensure harmonization and interoperability

[184] User requests for new features of the Kadi4Mat ELN will be collected through the GitLab web tool, where Kadi4Mat development is concentrated. GitLab allows end users and instance administrators to specify requirements for further development. A link to this GitLab form will be placed directly on the homepage of Kadi4Mat instances.
[185] The sub-activity output codes are linked to Annex 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.
[186] In the description of the KA, the planned activities are linked to the implemented outputs. The outputs are marked with a code in the format **P_x**.
[187] Information documenting the history of the described object (e.g. datasets) and related activities, which contains information about the origin or source of the described object, about any changes that have occurred since its creation, and who has taken charge of it since its creation.
[188] World Wide Web Consortium.
[189] Common Provenance Model.
[190] International Organization for Standardization.

with international activities and current developments at European (e.g. the EvolveBBMRI or UNCAN-Connect projects) and global level.

Provenance is an integral part of fulfilling the Reusable component of FAIR principles (the R in FAIR: "*R1.2: (Meta) data are associated with detailed provenance*"[191]) and applies to all datasets. In the activity, we strive for the development and subsequent pilot implementation of tools and services enabling the traceability of objects (*traceability*), primarily to support the assessment of the quality (suitability for use) of data stored in repositories and to support interoperability in an international context. The Provenance sub-activity will prepare materials and SW tools for the standardization of the preservation of provenance information in accordance with the ISO standard.[192]

It is now also possible to address the issue of Provenance within the NRP due to the fact that the underlying data model (CPM) for the representation of provenance is already mature enough for implementation within the NDI. Currently, the underlying data model is stable, and at the same time several prototypes for the use of the data model and standard have already been implemented, i.e. current developments permit the resolution of the issue.

The transfer of datasets and their linking between repositories requires:

- the existence of a methodology for working with provenance within the National Repository Platform in order to harmonize it;
- the implementation of SW to enable harmonized work with provenance (in this case according to ISO 23494) within the NRP and to ensure compatibility with international activities.

We will meet these two requirements as part one of outputs of the Provenance sub-activity and thus provide support for their primary users – developers, methodologists and repository managers, who can then integrate the implemented SW or follow the methodology for provenance management.

Provenance outputs will also be used in other KA projects, and in the following aspects in particular:

1. uniform approach to working with provenance within the NRP, which will allow the use of standard tools for its generation, storage and access;
2. alignment with international standards for working with provenance, in order to harmonize with developments at European or global level;
3. use of the current state-of-the-art in the field of provenance, which may create a background for future research in the area.

Specifically, the outputs will be at a minimum in the following key activities:

In the TKA SENSI, where a methodology for working with metadata (license and ethical/legal) will be prepared. As provenance can be perceived as a specific subset of metadata, this methodology should be consistent with the methodology for working with provenance. A new repository will be implemented in TKA DM4AI which will utilize SW tools to manage data created within the Provenance sub-activity. In the TKAs MATECH and B/H/F, the methodological foundations of the Provenance sub-activity will be used in order to integrate the new DANTE^c domain repository, or the Multiomic Human Data repository) with the current standards for provenance. New software, which includes the implementation of provenance in accordance with the methodology for working with provenance, will be

[191] https://www.go-fair.org/fair-principles/r1-2-metadata-associated-detailed-provenance/.

[192] Different domains deal with provenance in different ways. For example: ELNs are an example of tools that retain provenances, but not in a standardized form.

integrated in the ELN part of the PKA.

As part of the sub-activity, we will design and develop architecture for the provenance storage and management system. This software will enable: 1/ storage, access and versioning of document provenance[193], 2/ generation of metadata for document provenance according to the above standards, 3/ stored documents and related metadata will be identifiable by persistent identifiers for long-term preservation. These persistent identifiers will be stored as part of the metadata record and, after authorization, will be matched to the content of the stored document provenance.

The implemented system will also provide an API for communication with SW, and will be deployable in multiple instances and integrable with repository systems to allow searching for the predecessors of objects stored within repositories (known as backward traceability) and, if required, also outside them. The system integration itself is included in the TKA (see text above). The system will also allow access control at the level of the entire implemented document and will allow users to check the integrity and authenticity of stored document provenance using hashes and digital signatures. **P_1**

The implementation of the system for storing and administering the harmonised provenance will utilize a software library will be used; this will be implemented separately due to its usability in other systems. The software library will allow the:

- representation of the structures defined in the CPM;
- generation of document provenance using templates (templating mechanism) in accordance with the CPM;
- basic syntactic and semantic validation of the provenance of documents, i.e. verification that the provenance provided in the document's provenance is represented in accordance with the standards;
- searching for predecessors of objects stored in repositories. **P_2**

The implemented SW will be supplemented with a document containing a list of recommendations regarding the management of harmonised provenance in the NDI so that the management of provenance is in line with current and good practice and available standards related to the AI Act. A methodological document is being prepared in order to facilitate the integration of the software into repository systems and to harmonize work with provenance within the NDI. **P_3**

We plan to attend a conference (primarily Provenance Week) in order to keep up-to-date and maintain international contacts. Due to the potential international application of PKA provenance outputs, we also plan to present the outputs at an international conference (primarily the Extended Semantic Web Conference or Provenance Week).

ACTIVITIES:

- development of a system for storing and administering provenance; **P_1**
- development of a software library for work with provenance according to CPM; **P_2**
- design, pilot verification and final version of the methodology for provenance; **P_3**
- implementation of SW tool prototypes in 03/2027;
- draft methodology for provenance in 06/2027.

[193] Provenance is technically stored in a standardized machine-processable document (e.g. PROV-JSON).

SUB-ACTIVITY OUTPUT CODES[194]

P_1	SW for provenance storage and administration
P_2	SW library for work with provenance according to CPM
P_3	Methodology for provenance

[194] The sub-activity output codes are linked to Annex no. 3 Schedule of Key Activities and chap. 11 Main Planned Outputs/Products of the Feasibility Study.

MAIN PLANNED OUTPUTS

#ID	Name of the output	Output Type	Link to KA	Link to product in indicator 215 012 / 215 102	Brief description of the output	Method of submitting the output	Linking of the output to the output of another OP JAC project	Date of achievement of output[195]
B_1	National Omics Repository – Czech Omics Node (OmiCZ)	new repository	KA2	Repository set for TKA B/H/F	Repository for the secure storage and administration of sensitive genomic data, including: metadata model, integrated AAI, ProducerUI interface for data and metadata management, system for secure workflow and Galaxy execution in TRE, connection to FEGA, Beacon and ClinData. Incl. documentation. CF. ALSO PARTIAL OUTPUTS **B_1.1–B_1.6** KA 2.1.	web link to the repository, incl. technical documentation and user manual	YES – link to output(s) of the NRP	XII.28
B_2	Structural and simulation data repository (BioSimCZ)	new repository	KA2	Repository set for TKA B/H/F	A repository focused on structural simulation data, which will utilize a metadata model for simulation structural data as well as metadata models for selected aspects of biomacromolecules, and SW tools for extending the functions of the repository and documentation. CF. ALSO PARTIAL OUTPUTS **B_2.1–B_2.5** KA 2.2.	web link to the repository, incl. technical documentation and user manual	YES – link to output(s) of the NRP	XII.28
B_3	Repository for human and animal image and physiological	new repository	KA2	Repository set for TKA B/H/F	A repository for human and animal image and physiological multimodal data used	web link to the repository, incl.	YES – link to output(s) of the NRP	XII.28

[195] The deadline for achieving the output is the date of completion of the physical implementation of the project. However, it is appropriate to set appropriate deadlines in relation to the schedule of key activities, i.e. during the implementation of the project – depending on the nature of the output.

	multimodal data (Imaging Repository)				for the secure storage, administration and sharing of data. Incl. metadata model, SW tools for extending repository functions and documentation. CF. ALSO PARTIAL OUTPUTS **B_3.1**–**B_3.6** KA 2.3	technical documentation and user manual		
B_4	Chemical Biology Data Repository	new repository	KA2	Repository set for TKA B/H/F	A repository enabling the upload of selected chemical biology data, including their representation using RDF, using selected ontologies and effective search using the SPARQL language. CF. ALSO PARTIAL OUTPUTS **B_4.1**–B_**4.4** KA 2.4.	web link to the repository, incl. technical documentation and user manual	YES – link to output(s) of the NRP	XII.28
B_5	Linking of the FAIR repository with the development and training of AI models for MS	NDI development tools	KA2	Tools and services for the development of the NDI for TKA B/H/F	An interconnected environment for the development, training and validation of machine learning (AI) models focused primarily on the analysis of mass spectra.	software stored on Github	YES – complementarity of NRP	XII.28
B_6	ClinData Repository	new repository	KA2	Repository set for TKA B/H/F	A repository enabling the secure storage, administration and processing of data on people and their connection to other relevant repositories. CF. ALSO PARTIAL OUTPUTS **B_6.1**–**B_6.4** KA 2.5.	web link to the repository, incl. technical documentation and user manual	YES – link to output(s) of the NRP	XII.28
B_7	Methodology FAIRification of data from completed regulated clinical trials	methodology/standard	KA2	Set of methodologies for TKA B/H/F	Uniform methodological procedures enabling compliance with ECRIN standards and increasing the quality and availability of data, as well as the integration of data between domain-specific repositories.	link to the published electronic document	NO	XII.28
B_8	Adaptation of the ECRIN metadata model for the methodology for the FAIRification of data from completed regulated clinical trials.	methodology/standard	KA2	Set of methodologies for TKA B/H/F	A metadata model for data from completed regulated clinical trials, based on ECRIN standards and adapted to the needs of a medical data repository.	link to the published electronic document	NO	XII.28

B_9	Biological Imaging Data Repository Tools	NDI development tools	KA2	Tools and services for the development of the NDI for TKA B/H/F	Aggregated set of domain-specific tools for the Imaging Repository.	publicly available website that can access the tool	YES – complementarity of NRP	XII.28
M_1	DANTE[c] repository	new repository	KA3	TKA MATECH repository	DANTE[c] repository, with five communities and five collections encompassing metadata profiles, analyses, SW extending repository functions, and methodologies. CF. ALSO PARTIAL OUTPUTS **M_1.1–M_1.6** KA 3.1.	web link to the repository, incl. technical documentation and user manual	YES – link to output(s) of the NRP, EOSC CZ	XII.28
M_2	FAIR Implementation Profiles for communities in the DANTE[c] repository	methodology/standard	KA3	TKA MATECH methodology	Definition of standards for MATECH subdivisions, with a focus on collections in the DANTE[c] repository.	link to the published electronic document	YES – link to output(s) of the NRP	XII.28
M_3	Integration of the Kadi4Mat ELN with the DANTE[c] repository	services for the development of NDI	KA3	Tools and services for the development of the NDI for TKA MATECH	Establishment of processes in the Kadi4Mat ELN for the direct collection of data and metadata into collections of the DANTE[c] repository. Modification of the repository for receiving data from instances of the Kadi4Mat ELN.	publication on the direct sending of records from the ELN to the repository	NO	VI.27
M_4	Integration of the eLabFTW ELN with the DANTE[c] repository	services for the development of NDI	KA3	Tools and services for the development of the NDI for TKA MATECH	Establishment of processes in the eLabFTW ELN for the direct collection of data and metadata into collections of the DANTE[c] repository. Modification of the repository for receiving data from instances of the eLabFTW ELN.	publication on the direct sending of records from the ELN to the repository	NO	XII.28
M_5	SW tool for managing users and devices and their data outputs in research infrastructures	NDI development tools	KA3	Tools and services for the development of	Open-source SW tool (package) for the administration of users and devices and their data outputs for research	link to the source code of the developed open-source software	NO	XII.28

				the NDI for TKA MATECH	infrastructures with connection to the DANTE[c] repository and user documentation.	in the public Git repository		
M_6	integration of SW tools for managing users, devices and their data outputs in research infrastructures with the DANTE[c] repository	services for the development of NDI	KA3	Tools and services for the development of the NDI for TKA MATECH	Establishment of processes in the SW tool for data management in research infrastructures for the direct collection of data and metadata into collections of the DANTE[c] repository.	publication on the direct sending of records from the ELN to the repository	NO	XII.28
AI_1	DM4AI Repository	new repository	KA4	TKA DM4AI repository	A repository over the NRP infrastructure, assuming the use of NRP data repositories located near a large computing infrastructure, such as e-Infra CZ computing clusters, providing the user with the ability to store trained AI models, pipelines and datasets, and their metadata. CF. ALSO PARTIAL OUTPUTS **AI_1.1** – **AI_1.10** KA 4.1.	web link to the repository, incl. technical documentation and user manual	YES – link to output(s) of the NRP, EOSC CZ, CARDS	XII.28
AI_2	SW defining the environment for the inference of generative AI models stored in the data repository	NDI development tools	KA4	NDI development tools for TKA DM4AI	Environment for the inference of generative AI models stored in the data repository and interaction with them, with an emphasis on the efficiency of inference for different types and sizes of models.	at least one instance of the Clarin-DSpace repository system that will have an integrated user interface for communication with generative models	NO	XII.28
S_1	connection of the CSDA repository to the NRP	upgraded repository	KA5	TKA SOC repository set	Integration of expanded Dataverse system into the NRP, including integration of data collections. Contains metadata	web link to the repository, incl. identification of	YES – link to output(s) of the NRP	XII.28

	and its development and operation				models, data collections and includes analyses and methodologies. CF. ALSO PARTIAL OUTPUTS **S_1.1**–**S_1.10** KA 5.1.	upgraded elements		
S_2	New sensitive data repository for the social sciences	new repository	KA5	TKA SOC repository set	Building of a new repository for sensitive data from social science research using NRP resources, including a methodology and a pilot collection of sensitive data. CF. ALSO PARTIAL OUTPUTS **S_2.1**–**S_2.2** KA 5.2.	web link to the repository, incl. technical documentation and user manual	YES – link to output(s) of the NRP	XII.28
S_3	Upgraded DataHub repository	upgraded repository	KA5	TKA SOC repository set	Upgrade of the DataHub and the Map and Data Centre through the introduction of links to the NRP/NMD. CF. ALSO PARTIAL OUTPUTS **S_3.1**–**S_3.3** KA 5.3.	web link to the repository, incl. identification of upgraded elements	YES – link to output(s) of the NRP	XII.28
S_4	Data visualization tool	NDI development tools	KA5	NDI Development Tools for TKA SOC	Web-based visualization software tool for social science data from sample surveys showing the development of the attitudes, opinions and beliefs of the target population over time.	publicly available website that can access the tool	No	XII.28
S_5	Atlasobyvatelstva.cz platform for specialized maps	services for the development of NDI	KA5	NDI development service kit for TKA SOC	Integration of the Atlastobyvatelstva.cz platform for specialized maps into DataHub data services.	platform available as part of DataHub on the Web	NO	XII.28
S_6	Demographic database integrated into DataHub data services	services for the development of NDI	KA5	NDI development service kit for TKA SOC	A database focused on detailed demographic data of the Czech Republic at various geographical levels and timelines.	database available as part of DataHub on the Web	No	XII.28
S_7	Integration of external data sources	services for the development of NDI	KA5	NDI development service kit for TKA SOC	Performance of a landscape analysis focused on significant data producers and data sources. Data sources whose data cannot or should not be placed directly in	link to website	NO	XII.28

					the repository platform will be selected and made available.			
S_8	Analysis of sharing and practical use of data from international research infrastructures.	analysis	KA5	NO	Development of an analysis that will focus on ways of sharing and the practical use of data from international research infrastructures such as ISSP, EVS, SHARE ERIC, ESS ERIC and GGP.	link to the published electronic document	NO	XII.28
S_9	Communication strategy and dissemination of project outputs within the professional community in the social sciences	communication strategy	KA5	NO	Targeted communication strategy to increase awareness of data services and to improve the culture of data sharing and management in the Czech social sciences through a range of dissemination activities.	link to the published electronic document	NO	XII.28
S_10	Information centre on issues affecting new forms of data and the use of AI in research into data services	services for the development of NDI	KA5	NDI development service kit for TKA SOC	Sets of information materials, resources and procedures that will be included in the structure of CSDA services.	link to the published electronic document	No	XII.28
F_1	Creation and operation of a repository for the Physics domain	new repository	KA6	TKA PHYSICS repository	PHYSICS cluster repository with five communities, including metadata profiles. CF. ALSO PARTIAL OUTPUTS **F_1.1**–**F_1.6** KA 6.1.	web link to the repository, incl. identification of upgraded elements	YES – link to output(s) of the NRP	XII.28
F_2	Tool for the automated generation of metadata	NDI development tools	KA6	NDI development toolkit for TKA PHYSICS	Tool for the automated creation of metadata from measurements of physical quantities and their interconnection with the Electronic Laboratory Notebook (ELN).	publicly available website that can access the tool	NO	XII.28
F_3	Tool for the visual display of data files	NDI development tools	KA6	NDI development toolkit for TKA PHYSICS	A tool for the direct visual display of data files of measurements of physical quantities in a web browser.	publicly available website that can access the tool	YES – link to output(s) of the NRP	XII.28

H_1	LINDAT/CLARIAH-CZ repository – involvement of CNC infrastructure language corpora as a separate collection of the LINDAT/CLARIAH repository	upgraded repository	KA7	Set of repositories for TKA HUMA	Extension of the LINDAT/CLARIAH-CZ repository with a new Czech National Corpus community; its content will be the metadata of all publicly available corpora of the CNC. CF. ALSO PARTIAL OUTPUTS **H_1.1**–**H_1.9** KA 7.1.	web link to the repository, incl. identification of upgraded elements	YES – link to output(s) of the NRP	XII.28
H_2	Digitalia MUNI ARTS repository	upgraded repository	KA7	Set of repositories for TKA HUMA	A repository transforming research data collections in individual platforms into AIP packages in a specified data format and metadata structure. CF. ALSO PARTIAL OUTPUTS **H_2.1**–**H_2.4** KA 7.2.	web link to the repository, incl. identification of upgraded elements	YES – link to output(s) of the NRP	X.26
H_3	ArchaeoVault repository	upgraded repository	KA7	Set of repositories for TKA HUMA	The ArchaeoVault repository transforms research data collections in individual platforms into AIP packages in a specified data format and metadata structure and stores them in the NRP repository. Incl. analysis of best practice and data processing methodology. CF. ALSO PARTIAL OUTPUTS **H_3.1**–**H_3.6** KA 7.3.	web link to the repository, incl. identification of upgraded elements	YES – link to output(s) of the NRP	V.28
H_4	Repository for Bibliographical Data	new repository	KA7	Set of repositories for TKA HUMA	A new repository for the collection, storage and accessing of bibliographical data, including metadata models, workflows and methodologies for working with data. CF. ALSO PARTIAL OUTPUTS **H_4.1**–**H_4.9** KA 7.4.	web link to the repository, incl. technical documentation and user manual	YES – link to output(s) of the NRP	XII.28
H_5	A set of superstructure tools for the development of TKA HUMA repositories	NDI development tools	KA7	NDI development toolkit for TKA for HUMA	A set of additional tools for data processing and visualization, providing greater ease of working with the content of specific repositories. CF. ALSO PARTIAL OUTPUTS **H_5.1**–**H_5.4** KA 7.5.	publicly available website that can access the tool	YES – link to output(s) of the NRP	XII.28

E_1	Upgraded GENASIS repository	upgraded repository	KA8	Set of repositories for TKA ENVIRO	An upgraded repository for the presentation of data on chemical exposure of environmental and human tissue matrices, enabling the integration of various data, including a metadata model, metadata sets and FAIRified datasets. CF. ALSO PARTIAL OUTPUTS **E_1.1**–**E_1.4** KA 8.1.	web link to the repository, incl. identification of upgraded elements	YES – link to output(s) of the NRP	XII.28
E_2	Repository for storing data from non-targeted mass spectroscopic analyses for human exposome assessment	new repository	KA8	Set of repositories for TKA ENVIRO	A new repository for primary mass spectroscopic data from non-targeted metabolomics and exposomics for human samples, including a metadata model, metadata sets and FAIRified datasets. CF. ALSO PARTIAL OUTPUTS **E_2.1**–**E_2.3** KA 8.2.	web link to the repository, incl. technical documentation and user manual	YES – link to output(s) of the NRP	XII.28
E_3	Repository for the storage of toxicological and ecotoxicological data	new repository	KA8	Set of repositories for TKA ENVIRO	Repository for the storage of toxicological and ecotoxicological data, incl. needs analysis, incl. metadata model and FAIRified datasets. CF. ALSO PARTIAL OUTPUTS **E_3.1**–**E_3.4** KA 8.3.	web link to the repository, incl. technical documentation and user manual	YES – link to output(s) of the NRP	XII.28
E_4	New repository for linking geocoded data from different domains	new repository	KA8	Set of repositories for TKA ENVIRO	A new repository that facilitates the interconnection of geocoded data from different areas in order to study health risks. Incl. needs analysis to serve as the basis for the creation of the design and metadata model for this repository. CF. ALSO PARTIAL OUTPUTS **E_4.1**–**E_4.2** KA 8.4.	web link to the repository, incl. technical documentation and user manual	YES – link to output(s) of the NRP	XII.28
E_5	New repository for reference image data of living plants and plant communities	new repository	KA8	Set of repositories for TKA ENVIRO	New repository for reference image data of living plants and plant communities, incl. metadata model, metadata sets, FAIRified datasets and user methodology.	web link to the repository, incl. technical	YES – link to output(s) of the NRP	XII.28

					CF. ALSO PARTIAL OUTPUTS **E_5.1**–**E_5.4** KA 8.5.	documentation and user manual		
E_6	New repository for the genetic biomonitoring and genetic data of wild organisms	new repository	KA8	Set of repositories for TKA ENVIRO	New repository for genetic biomonitoring and genetic data of wild organisms, including a metadata model, metadata sets and FAIRified datasets and a methodology for users. CF. ALSO PARTIAL OUTPUTS **E_6.1**–**E_6.4** KA 8.6.	web link to the repository, incl. technical documentation and user manual	YES – link to output(s) of the NRP	XII.28
E_7	New repository for zoological collections	new repository	KA8	Set of repositories for TKA ENVIRO	New repository for zoological collections. Extension of the thematic cluster of the Repository for Biodiversity Data with a new section for non-genetic zoological collections (taxidermy and alcohol-preserved preparations, skeletons) and its connection with the section for herbarium collections (in preparation). CF. ALSO PARTIAL OUTPUTS **E_7.1**–**E_7.4** KA 8.7.	web link to the repository, incl. technical documentation and user manual	YES – link to output(s) of the NRP	XII.28
E_8	New bioinformatics tools for the analysis of mass spectroscopic data from non-targeted metabolomic and exposome analyses	NDI development tools	KA8	NDI development toolkit for TKA ENVIRO	Standardized and documented tools for high-resolution mass spectroscopy data mining for the qualitative and quantitative analysis of a wide range of exogenous and endogenous substances.	publicly available website that can access the tool	YES – link to output(s) of the NRP	XII.28
E_9	AI tools for the image analysis of biological objects and the automatic reading of herbarium labels.	NDI development tools	KA8	NDI development toolkit for TKA ENVIRO	Deployment and modification of the AI model for image analysis of herbarium items (e.g. HESPI https://rbturnbull.github.io/hespi/), integration of LLM models for automating the transcription of metadata from herbarium labels into the repository workflow.	publicly available website that can access the tool	YES – link to output(s) of the NRP	XII.28

SE_1	Analysis of good practice for the sharing of sensitive data from collaborative research	analysis	KA 9	NO	A document summarizing good practice for the sharing sensitive data from collaborative research.	link to the published electronic document	NO	III.27
SE_2	Methodology FAIRification of Sensitive Data in NDI	methodology/standard	KA 9	Methodology for TKA SENSI	Comprehensive methodology for the comprehensive management of sensitive data in NDI.	link to the published electronic document	YES – link to output(s) of the NRP	VI.28
SE_3	Software module for checking and ensuring the quality of repository data	NDI development tools	KA 9	NDI development tools for TKA SENSI	Software module for checking and ensuring the quality of data in NRP repositories developed in the ClinData eCRF system.	publicly available website for accessing the tool, including documentation	YES – complementarity of NRP	XII.28
SE_4	Service for a controlled, secure process for receiving/dispensing data to/from the environment for processing sensitive data	Services for the development of the NDI	KA 9	NDI development service kit for TKA SENSI	Implementation of a secure controlled process for receiving/dispensing data to/from the environment for processing sensitive data.	technical implementation of the service available to researchers in the Czech Republic, including documentation	YES – link to output(s) of the NRP	XII.28
SE_5	End-user interface for the interactive/batch processing of sensitive data in a secure environment	Services for the development of the NDI	KA 9	NDI development service kit for TKA SENSI	Identification of suitable interfaces for the interactive and batch processing of sensitive data in a secure environment – these will be implemented.	technical implementation of the service available to researchers in the Czech Republic, including documentation	YES – link to output(s) of the NRP	XII.28
SE_6	Service for the authorization of batch processing of sensitive data	Services for the development of the NDI	KA 9	NDI development service kit for TKA SENSI	Establishment and implementation of a mechanism that will allow authorization of requests for the batch processing of	technical implementation of the service available to researchers in	NO	XII.28

					sensitive data in a TRE-type environment.	the Czech Republic, including documentation		
SE_7	Service for the evaluation of analytical workflows and classification of the sensitivity of their outputs	Services for the development of the NDI	KA 9	NDI development service kit for TKA SENSI	The service will provide a systematic evaluation of the sensitivity of the outputs of analytical workflows.	technical implementation of the service available to researchers in the Czech Republic, including documentation	YES – complementarity of NRP	XII.28
LLM_1	Technical report: collection of requirements and design of architecture for the use of large language models	technical report	KA 10	NO	Technical report with the conclusions of the analysis of the requirements for the virtual assistant technology in terms of social, legal, psychological and ethical aspects and the created design of architecture and processes for the machine processing of user inputs.	link to the published electronic document	NO	II.26
LLM_2	Resulting version of the virtual assistant system for deployment in the helpdesk environment	NDI development tools	KA 10	Set of tools for the development of NDI per PKA	A system with technical documentation and documentation for operators, which will be piloted and operated for user-researchers and L2/L3 helpdesk staff.	link to the source code of the developed open-source software in the public git repository	NO	XII.28
KB_1	Security monitoring system requirements	technical report	KA 10	NO	Technical report describing the conclusions of the analysis of the requirements for the security monitoring system and the architecture design created, including a description of the individual components of the system.	link to the published electronic document	NO	II.26

KB_2	Recommended practices for the Open Science environment for effective defence against cyber attacks	technical report	KA 10	Set of methodologies per PKA	Recommended practices for securing the specific environment of the NRP. Set of proven methods and strategies for effective defence against cyber-attacks.	link to the published electronic document	NO	XII.26
KB_3	Proactive security monitoring toolkit	NDI development tools	KA 10	Set of tools for the development of NDI per PKA	Proactive security monitoring toolkit	link to the source code of the developed open-source software in the public git repository	YES – complementarity of NRP	XII.27
KB_4	Security threat detection tools and reactive security support	NDI development tools	KA 10	Set of tools for the development of NDI per PKA	Set of tools for detecting security threats and providing information to support reactive security.	link to the source code of the developed open-source software in the public git repository	YES – complementarity of NRP	XII.28
KB_5	Documentation of the developed ecosystem and pilot deployment report	technical report	KA 10	Set of methodologies per PKA	Documentation of the developed ecosystem, providing a detailed overview of the architecture, functionalities and integration of individual developed components of the system.	link to the published electronic document	YES – complementarity of NRP	XII.28
C_1	Set of recommendations and methodology, especially in the field of artificial intelligence and data processing	technical report	KA 10	Set of methodologies per PKA	Document focusing on the issue of using the LLM from the point of view of law. Summarizes the state and procedure for implementing technological security measures and user support using AI methods.	link to the published electronic document	YES – complementarity of NRP	XII.28
ELN_1	Integration of eLabFTW ELN with NRP repositories	NDI development tools	KA 10	Set of tools for the	An implemented tool that allows the user to publish data and metadata from records in the Kadi4Mat ELN to various	technical implementation of the service available	YES – link to output(s) of the NRP	VI.27

				development of NDI per PKA	domain data repositories and to the catch-all NRP repository in the NDI.	to researchers in the Czech Republic, including documentation		
ELN_2	Integration of the Kadi4Mat ELN into NRP repositories	NDI development tools	KA 10	Set of tools for the development of NDI per PKA	An implemented tool that allows the user to publish data and metadata from records in the eLabFTW ELN to various domain data repositories and to the catch-all NRP repository in the NDI.	technical implementation of the service available to researchers in the Czech Republic, including documentation	YES – link to output(s) of the NRP	XII.28
ELN_3	Deployment methodology for the Kadi4Mat and eLabFTW ELNs	methodology/standard	KA 10	Set of methodologies per PKA	Methodology for deploying the Kadi4Mat and eLabFTW ELNs in scientific organizations that will create and administer their own instance using NDI resources.	link to the published electronic document	YES – link to output(s) of the NRP	XII.27
ELN_4	Deployment of catch-all ELN instances	NDI development tools	KA 10	Set of tools for the development of NDI per PKA	1 instance of Kadi4Mat ELN, 1 instance of eLabFTW ELN.	technical implementation of the service available to researchers in the Czech Republic, including documentation	YES – link to output(s) of the NRP	XII.26
ELN_5	Upgraded ELN Kadi4Mat	NDI development tools	KA 10	Set of tools for the development of NDI per PKA	Interconnection of different instances of the Kadi4Mat ELN, which will allow users to interactively share data between instances after authorization through mutual direct access without export and import.	Source code publicly available along with the history of its development	NO	XII.28
P_1	SW for provenance storage and administration	NDI development tools	KA 10	Set of tools for the	SW for checking the integrity and authenticity of stored document provenance using hashes and digital signatures.	publicly available website that can access the tool	YES – complementarity of NRP	VI.28

				development of NDI per PKA				
P_2	SW library for work with provenance according to CPM	NDI development tools	KA 10	Set of tools for the development of NDI per PKA	The SW library will allow the representation of structures defined in the CPM; generation of document provenance using templates, etc.	publicly available website that can access the tool	YES – complementarity of NRP	VI.28
P_3	Methodology for provenance	methodology/standard	KA 10	Set of methodologies per PKA	Document containing a list of recommendations for the management of harmonised provenance in the NDI so that the provenance report is in line with current best practice and available standards related to the AI Act.	link to the published electronic document	YES – complementarity of NRP	XII.28
K_2	Training materials kit for OS II	training material	KA 10	OS II educational materials	The set of educational materials encompasses: **AI_1.11** (KA 4) Educational materials/tutorials; **S_10.1** (KA 5) Materials for the data management education and methodology of social science research **S_10.2** (KA 5) Creation of guidelines for data management education and methodology of social science research **F_4.1** (KA 6) Series of three interactive e-learning courses for PHYSICS **F_4.2** (KA 6) Educational materials for four PHYSICS workshops; **H_1.9** (KA 7) FAIR data in the deployment and operation of the DSpace repository instance within EOSC CZ; **H_4.10** (KA 7) Educational materials for work with the Repository for	link to the published electronic document	YES – basis for training events organized by the EOSC CZ Training Centre	XII.28

					Bibliographical Data; **E_1.4** (KA 8) Standard operating procedures and tutorials for the creation and validation of FAIR-validified datasets; SE_2.1 (KA 9) Educational materials connected to the FAIR-sensitive data methodology.			
K_1	OS II Conference	Conference	KA 10	NO	Four professional domain conferences that will provide a platform for sharing project outputs and networking for the professional community.	Conference programme, presentations	YES – cooperation with the EOSC CZ Training Centre	XII.28