This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

IDEA4RC

IDEA4RC

Intelligent ecosystem to improve the governance, the sharing, and the re-use of health data for rare cancers

Deliverable 9.1

Pilot deployment plan

31 August 2024

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Distribution List

Organization	Name of recipients
1 - Coord INT	A. Trama, P. Casali, L. Buratti, P. Baili, J. Fleming, L. Licitra, E. Martinelli, G. Scoazec
2 - UDEU	A. Almeida, , U. Zulaika Zurimendi, N.. Kalocsay
3 - MME	F. Mercalli, S. Copelli, M. Vitali
4 - UPM	E. Gaeta, G. Fico, L. Lopez, I. Alonso, A. Alonso, L. Hernandez, C. Vera
5 - HL7	G. Cangioli, C. Chronaki
6 - ECCP	S. Ziegler, S. Schiffner, V. Tsiompanidou
7 - ENG	P. Zampognaro, A. Sperlea, E. Mancuso, M. Melideo, F. Saccà, V. Falanga, M. Rosa
8 - CERTH	K. Votis, A. Triantafyllidis, N. Laloumis, I Drympeta
9 - UU	S. van Hees, Wouter Boon, E. Moors, M. Kahn-Parker, C. Egher
10 - DICOR	C. Lombardo, G. Pesce, G Ciliberto, A. Tonon,
10° - ACC (Affiliated)	D. De Persis, P. De Paoli, G. Piaggio, M. Pallocca. A. De Nicolo
11 - FBK	A. Lavelli, S. Poggianella, O. Mayora, A.M. Dallaserra
12 – IKNL	E. Bosma. G. Geleijnse, A. Van Gestel, F. Martin, E. Mezei
13 - CLB	A. Sans, M. Brahmi, A. Pons, J-Y Blay, H. Crochet, J. Olaz, J. Bollard, C. Chemin-Airiau, H. Crochet
14 - APHP	B. Baujat, E. Koffi
15 - FJD	J Martin-Broto, N. Hindi, M. Martin Ruiz, A. Montero Manso, C. Roldàn Mogìo, D. Da Silva, A. Herrero, B. Barrios
16 - VGR	Magnus Kjellberg, L. De Verier, A. Muth
17 - MSCI	I. Lugowska, D. Kielczewska, M.Rosinska, A KAwecki , A., P. Rutkowski
18 - MUH	R. Knopp, A. Sediva, K. Kopeckova, A. Nohejlova Medkova, M. Vorisek
19 - OUS	S. Larønningen, J. Nygård, M. Sending, O. Zaikova
20 - MMCI	J. Halamkova, I. Mladenkova, l. Tomastik, V. Novacek, T. Kazda, I. Mladenkova, O. Sapožnikov, V. Novacek
21 - CLN	R. Szmuc, J. Poleszczuk, R. Lugowski
22 - FPNS	M. Barbeito Gomez, P. Parente, L. Carrajo Garcia, P. Ramos Vieiro
23 - TNO	E. Lazovik, L. Zilverberg, S. Dalmolen
24 - INF	ML Clementi, C. Sabelli
25 - UKE	S. Bauer, S. Lang, S. Mattheis, N. Midtank, M. Kim

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Revision History

Revision	Date of Issue	Author(s)	Brief Description of Change
v0.1	June 2024	UPM, MME	Table of contents
V0.5	July 2024	UPM	Chapter 2 and 3
V1	August 2024	UPM, MME	Integration of Chapter 4 and 5 Chapters 1 and 6 Table of abbreviations
V1	August 2024	INT	First revision
V2	02-09-2024	UPM	Revised version
		CLB	Final Peer-review

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Addressees of this document

This document is addressed to the whole IDEA4RC Consortium. It is an official deliverable for the project and shall be delivered at the European Commission and appointed experts.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

TABLE OF CONTENTS

1 Executive summary 9
2 Objective of the deployment 10
2.1 Needs for pilot data sharing and federated research 10
2.1.1 Research for rare cancer 10
2.1.2 Structuring unstructured data 10
2.1.3 Data preparation and ingestion 11
2.1.4 Privacy preserving environments for secondary use of data 11
2.1.5 Governance 12
2.1.6 Multimodal navigator 12
2.2 Technical deployment objectives 12
2.2.1 Research for rare cancer 13
2.2.2 Structuring unstructured data 13
2.2.3 Data preparation and ingestion 14
2.2.4 Privacy preserving environments for secondary use of data 14
2.2.5 Governance layer integration 15
2.2.6 Multimodal navigator integration 16
3 Deployment Plan 17
3.1 Phase 1: Proof of concept and real testing [Sep23, Aug24] 17
3.1.1 Understanding the data model [Sept23, Oct23] 18
3.1.2 Test deployment [Oct23, Nov23] 18
3.1.3 Demo data collection [Nov23, Jan24] 19
3.1.4 Demo capsule deployment [Nov24, Jan25] 20
3.1.5 ETL test [Dec24, Jun25] 21
3.1.6 Collecting structured data [Jan24, Aug24] 21
3.1.7 Production capsule deployment [Jan24, Jun24] 22
3.1.8 ETL on structured data [Jan24, Jun24] 22
3.1.9 Federated learning integration [Dec23, Aug24] 23
3.1.10 NLP algorithm definition [Sep23, Feb24] 23
3.1.11 NLP training [Jan24. Aug24] 24
3.1.12 Metadata services [Jan24, Aug24] 25
3.1.13 NLP data collection [Jan24, Jun24] 25
3.1.14 Governance [May24, Aug24] 26
3.1.15 Definition and design of the analytics and multimodal navigators [Nov23, Aug24] 26
3.1.16 Data mapping at CoEs [Dec23, Aug24] 27

3.1.17 Data quality [Dec23, Aug24] ... 27
3.1.18 Data annotation tool integration [Dec23, Apr24] ... 28
3.1.19 Milestone 1: Capsule version 1 up and running in the CoEs with no-personal data ... 28
3.2 Phase 2: Implementation and scaling [Sep24, Sep25] ... *29*
3.2.1 Fake data capsule deployment [Sept24, Oct24] ... 30
3.2.2 Preparation of production environment [Sept24, Feb25] ... 30
3.2.3 Production capsule deployment [Feb25, Mar25; Jul25, Sep25] ... 30
3.2.4 ETL engines [Sep24, Feb 25; Apr25, Aug25] ... 31
3.2.5 Collecting structured data [Sep24, Dec24] ... 31
3.2.6 Novel Federated Learning algorithm integration [Sep24, Jul25] ... 31
3.2.7 NLP data collection [Sept24, Dec24] ... 31
3.2.8 NLP training [Oct24, Jan25] ... 32
3.2.9 NLP integration [Nov24, Jan25; Apr25, Jul25] ... 32
3.2.10 Metadata services [Sept24, Jul25] ... 32
3.2.11 Governance integration [Sept24, Jul25] ... 32
3.2.12 Integration of augmented analytics and multimodal navigator [Sep24, Jul25] ... 32
3.2.13 Data mapping at CoEs [Sept24, Jan25] ... 33
3.2.14 Data quality [Sep24, Jan25] ... 33
3.2.15 Data annotation tool integration [Sept24, Oct24] ... 33
3.2.16 Milestone 2: Capsule version 2 up and running in the CoEs with real data ... 33
3.2.17 Milestone 3: Capsule version 3 VA and governance integrated ... 35
4 Adaptable DevOps culture for Deployement, management, support and Execution ... 37
4.1 Management ... *37*
4.1.1 Roles and contact persons: actors' map and WG on Pilot Deployment ... 37
4.1.2 Communication ... 39
4.2 Support and monitoring ... *41*
4.2.1 Periodic and ad hoc meetings ... 41
4.2.2 Activity monitoring ... 43
4.2.3 Webinars ... 45
5 Adapting to local pilot specificities ... 47
6 Conclusions ... 55

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

LIST OF FIGURES

Figure 1 – Gantt phase 1 18
Figure 2 – Gantt phase 2 30
Figure 3 – Actors' map 38
Figure 4 – Slack channels 40
Figure 5 – Shared repository for the WG on Pilot Deployment, example excerpt 41
Figure 6 – Example of materials for a Monthly Pilot Meeting 42
Figure 7 – Excerpt of TODO List for WG on Pilot Deployment monitoring 43
Figure 8 – Screenshot form a Pilot Deployment Webinar 45
Figure 9 – Example of materials for a WG on Pilot Deployment Webinar 46

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Abbreviations and definitions

Abbreviation	Definition
AI	Artificial intelligence
CoEs	Centres of Excellence
DWH	Data Warehouses
EHDS	European Health Data Space
ETL	Extract, Transform, Load
MPC	multiparty computation
NER	named entity recognition
NLP	Natural Language Processing
PoI	Point of injection
UAT	User acceptance testing
UI	User Interface
VA	Virtual Assistant
VM	Virtual Machine
WP	Work Package
WG	Working Group

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

1 EXECUTIVE SUMMARY

This deliverable outlines a comprehensive deployment plan for the IDEA4RC pilot focused on rare cancer epidemiology. The primary goal is to enable secure, privacy-preserving environments that facilitate the secondary use of structured and unstructured data across several federated Centres of Excellence (CoEs). IDEA4RC deployment covers the implementation of advanced technologies such as Natural Language Processing (NLP) algorithms, federated learning, and augmented analytics within a multimodal navigator.

The deployment plan is structured into two main phases: Phase 1, which covers the proof of concept and initial testing, and Phase 2, focused on scaling-up and full implementation. Throughout these phases, three key milestones are identified. Milestone 1 involves the deployment of the first capsule version with no-personal data. Milestone 2 follows with the deployment of capsule version 2, incorporating real data within a capsule within the CoEs infrastructure. Milestone 3 culminates with the deployment of capsule version 3, which includes the full integration of value-added services and governance mechanisms. These milestones collectively demonstrate the project's progression toward a fully functional, secure, and efficient federated data processing ecosystem. Governance, data quality, and adaptable DevOps practices are emphasized to ensure successful execution.

This deliverable also addresses the specific needs of different pilot sites, ensuring flexibility and adaptability to local requirements. The deployment plan is designed to be collaborative and iterative, with continuous support and monitoring through structured communication channels, regular meetings, and webinars.

2 OBJECTIVE OF THE DEPLOYMENT

2.1 Needs for pilot data harmonization and federated research

The deployment of the IDEARC platform requires addressing several key objectives. In this context, five main needs have been identified as crucial for successful pilot privacy preserving data processing and federated research. These include curating and aggregating existing data on rare cancers for improving research by means of: structuring unstructured data; ensuring accurate and consistent data extraction, quality, preparation, and ingestion; creating privacy-preserving environments for secondary use of data; establishing robust and scalable governance frameworks; and developing multimodal navigators to facilitate efficient data retrieval and interpretation.

These efforts align closely with the goals of the European Health Data Space (EHDS)[1], which aims to create a unified framework for health data exchange across Europe. By leveraging the principles and infrastructure outlined in the EHDS, the IDEARC platform can ensure that data is handled securely and efficiently, promoting innovation and collaboration in healthcare research. The EHDS supports the creation of a connected and interoperable health data ecosystem, facilitating access to high-quality data while maintaining strict privacy standards. This synergy between IDEARC's objectives and the EHDS framework underscores the potential for significant advancements in medical research and patient care across Europe.

2.1.1 Research for rare cancer

The research for rare cancer epidemiology faces a significant challenge due to the limited availability of data on rare cancers and more in general on rare diseases. This scarcity of data necessitates a critical need for curating and aggregating the existing data. By systematically organizing and combining available information, researchers can enhance the quality and comprehensiveness of data, which is essential for advancing the understanding and treatment of these rare conditions.

2.1.2 Structuring unstructured data

One of the main challenges in researching rare cancers is that the relevant data is primarily hidden in unstructured free-text fields, such as pathology reports. This makes it essential to

[1] https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en

structure this data to make it usable for analysis and research. This is evident from the acceptance criteria mentioned in the article of Guergana & Co[2], which highlights the importance of extracting coded variables from free-text fields, such as pathology reports, date of diagnosis, or treatments, using NLP techniques.

The acceptance criteria in the document underscore the importance of extracting coded variables from these free-text fields using Natural Language Processing (NLP) techniques. By converting unstructured data into a structured format, researchers can improve data accessibility and enhance the overall quality of research on rare cancers.

2.1.3 Data preparation and ingestion

Ensuring accurate and consistent data from diverse sources and Centres of Excellence (CoEs) is crucial for reliable cancer research outcomes. The European Platform on Rare Disease Registration aims to make health data searchable, findable, and valuable by promoting the extended reuse and standardization of data sharing across hundreds of registries within different health systems in Europe. This highlights the need for an efficient system to transform and load data into analysis systems. Given the lack of a reference infrastructure architecture for establishing a reliable system for data preparation and ingestion is essential to achieve high-quality data and trustworthy research results.

2.1.4 Privacy preserving environments for secondary use of data

Cancer data is highly sensitive and must be protected to prevent misuse; however, sharing and making this data available is crucial for advancing cancer research. The need to protect sensitive cancer data while making it accessible for research purposes presents a significant challenge. This data often contains personal information that requires protection from unauthorized access, copying, modification, or removal. Nonetheless, ensuring that this data can be safely harmonized and processed is essential for advancing cancer research and improving patient outcomes. Establishing privacy-preserving environments is therefore critical to balance data protection with the need for scientific progress.

[2] Guergana K. Savova, Ioana Danciu, Folami Alamudun, Timothy Miller, Chen Lin, Danielle S. Bitterman, Georgia Tourassi, Jeremy L. Warner; Use of Natural Language Processing to Extract Clinical Cancer Phenotypes from Electronic Medical Records. Cancer Res 1 November 2019; 79 (21): 5463–5470. https://doi.org/10.1158/0008-5472.CAN-19-0579

2.1.5 Governance

Each CoE has its own governance structure with distinct policies, procedures, and standards, making it difficult to ensure privacy-preserving processing across federated data from multiple CoEs. According to the GDPR, the Data Controller, who determines the purposes and means of processing personal data, is a key stakeholder in this scenario. To address these challenges, it is essential to assess and implement the governance frameworks of each CoE, including the roles, responsibilities, and accountability mechanisms for data controllers, processors, and other entities involved in data sharing. This evaluation should include mechanisms for monitoring compliance and handling potential breaches or incidents. Additionally, it is crucial to ensure that individuals' rights are respected throughout the federated data processing, including the rights to access, rectify, restrict processing, or erase personal data. Addressing these governance issues requires a thorough analysis of the existing frameworks across CoEs and the implementation of measures to ensure accountability, transparency, and respect for individual rights.

2.1.6 Multimodal navigator

The fragmentation, complexity, and lack of interoperability of data across diverse sources, particularly in healthcare, pose significant challenges for efficient information retrieval and decision-making. Traditional search methods often struggle to find relevant data researchers are looking for, leading to inefficiencies. The need for a multimodal navigator arises from this problem, as it aims to integrate various data formats into a unified interface and automate their interpretation. This tool enhances accessibility, usability, and contextual understanding of the data, thereby improving efficiency and productivity in data-driven domains.

2.2 Technical deployment objectives

This section describes which technical solution will be implemented and deployed at each CoE to address each need described in the previous section. The solutions will be aligned as much as possible with the results of the existing initiatives related with EHDS especial with EHDS2pilot[3], TEHDAS2[4] and Quantum project[5].

[3] https://ehds2pilot.eu/
[4] https://tehdas.eu/
[5] https://quantumproject.eu/

2.2.1 Research for rare cancer

Curating and aggregating existing data on rare cancers will be addressed in IDEA4RC with several development and tools. First we are working in WP2 and WP4 on a common data model designed for addressing cancer research use cases in the 4 main area described in WP8:

- Description of the natural history of soft tissue sarcoma and head and neck cancers
- Evaluation of factors that influence prognosis and treatment response
- Assessment of the treatment effectiveness (systemic, radiotherapy, surgery, target therapy, immunotherapy and possible combinations)
- Quality of care (diagnostic and staging procedures, treatment strategies, follow-up etc.).

Data is then mapped on the most commlny used standards OMOP and FHIR (WP2 and WP4) that allows federated analysis in WP4. Data quality tools are also provided for curating and harmonizing data against selected benchmarks and make them available and discoverable in a virtual setting that grant ownership and privacy of the data (WP4).

Finally, the standardized curated data model is manipulated to address each of the specific research areas through the Vantage6 platform (T4.3) that enables privacy preserving data processing in each CoE.

All these components are expected to be integrated and deployed into the CoE capsules. CoE will be engaged in several activities for the integration of the tools and the definition of the data model.

2.2.2 Structuring unstructured data

To structure unstructured data, defining a curated and standardized data model aligned with the expected research outcomes is crucial, but not sufficient. We need to go a step further and structure the data hidden in free text. This requires the development and integration of NLP algorithms for named entity recognition (NER) in the CoEs capsules (WP5). These algorithms will enable us to extract relevant information from unstructured free-text fields, making it possible to perform federated privacy-preserving analysis across different capsules.

However, to provide high performance for these NLP algorithms, several interactions and agreements are expected to share existing data across several centres. This will allow us to better train the models, ensuring that they can accurately identify and extract relevant information from free text. By doing so, we can unlock the potential of unstructured data and gain a deeper understanding of rare cancers.

2.2.3 Data preparation and ingestion

In order to ensure FAIR data, two major and related activities are essential for both structured and unstructured data at each CoE: data preparation and data ingestion.

The data preparation phase involves all the necessary activities to prepare local data for the creation of the local capsule. Specifically, every CoE will be required to align their data to the IDEA4RC data model (WP3) and annotate existing unstructured data in line with the IDEA4RC data model and selected vocabularies. This annotation process is critical for training and fine-tuning the NLP models used for NER to the specificities and objectives of IDEA4RC (WP5).

By performing these activities locally at each CoE, we can ensure that data from diverse sources is transformed and can be ingested into analysis systems efficiently and consistently, enabling high-quality cancer research outcomes. The ingestion phase expects that data is formatted into standard, in IDEA4RC FHIR and OMOP will be supported, after that the capsule will be ready for performing privacy preserving data analysis.

2.2.4 Privacy preserving environments for secondary use of data

Imagine a scenario where several hospitals want to collectively analyse patient data without compromising individual privacy. In this setup, each hospital maintains its data within its secure environment. Instead of pooling all data into a central repository, which could pose significant privacy risks, federated data analysis enables these hospitals to collaborate on analysis tasks while ensuring each dataset remains confidential and under the control of its respective owner.

The technical mechanisms behind federated data analysis involve innovative approaches such as secure computation protocols. These protocols utilize advanced cryptographic techniques like secure multiparty computation (MPC) or homomorphic encryption. They allow computations to be performed on encrypted data or ensure that data remains private during collaborative analysis. For instance, hospitals can independently perform machine learning model training or statistical analyses on their data without sharing raw patient records.

Once local analyses are conducted, the results—such as aggregated statistics or model updates—are securely aggregated by a designated aggregator. This aggregator collects and combines the outcomes from all participating entities without accessing sensitive individual-level data. The aggregated insights can then be shared among participants for further refinement or decision-making purposes.

The benefits of federated data analysis are substantial. First and foremost, it upholds data privacy regulations and safeguards sensitive information. By distributing data analysis tasks across multiple entities, it minimizes the risk of data breaches or unauthorized access. Moreover, it promotes collaboration and knowledge sharing without compromising confidentiality, making it particularly valuable in sectors like healthcare, finance, and telecommunications.

Overall, federated data analysis represents a forward-thinking approach to leveraging data for collective insights while respecting individual privacy rights. As advancements in secure computation and federated learning frameworks continue, this approach is likely to become even more integral to data-driven collaboration across diverse domains.

Within IDEA4RC the federated approach will be implemented by means of the IDEA4RC capsules where the Vantage6 platform will be integrated. This platform is developed by IKNL that is the WP4 leader.

2.2.5 Governance layer integration

To create a governance layer that allows the heterogeneity of Every CoE in policies, procedures, and standards, which can lead to difficulties in ensuring consistent and compliant data sharing practices, IDEA4RC is implementing an additional platform that digitalizes the processes of legally allowing the use of data while respecting privacy-preserving agreements. This platform will consider the specificities of each CoE, integrating with the technical components designed to perform data analysis.

The current manual process requires a lot of interaction and delivery of documentation aligned with administrative and legal procedures that are different for each CoE and country. By digitalizing these processes, we can streamline the governance framework in place for each CoE, including roles, responsibilities, and accountability mechanisms for data controllers, processors, and other entities involved in the data sharing.

This evaluation should consider mechanisms for monitoring compliance and handling any potential breaches or incidents. Moreover, it is crucial to ensure that individuals' rights are respected and upheld throughout the data sharing process, including the right to access, rectify, restrict processing, or erase their personal data. This can be achieved by evaluating mechanisms for ensuring transparency and communication, such as providing clear and

concise privacy notices, informing individuals about the purposes, risks, and rights related to the data sharing.

By implementing this platform, IDEA4RC aims to provide a solution that addresses the complexities of governance frameworks across multiple CoEs, ensuring accountability, transparency, and respect for individual rights.

2.2.6 Multimodal navigator integration

To address the problem of fragmentation, complexity, and lack of usability of User Interfaces (UIs) across diverse data sources, especially evident in fields like healthcare, IDEA4RC is developing a virtual assistant called multimodal navigator (WP6). This navigator is designed to integrate disparate data types into a unified interface, automate interpretation, and enhance accessibility, usability, and contextual understanding. By doing so, they aim to improve efficiency and productivity in data-driven domains.

The multimodal navigator will play a crucial role in enabling researchers of WP8 to perform federated data analysis, from the discoverability of the data to the sharing of the results of the data analysis. To achieve this, parts of this navigator need to be integrated into the CoE capsules, ensuring seamless interaction and leveraging the digitalized governance framework established by IDEA4RC.

By integrating the multimodal navigator with the CoE capsules, researchers will be able to efficiently retrieve relevant information, make informed decisions, and accelerate the data analysis process. This will ultimately lead to improved outcomes in healthcare and other data-driven domains.

3 DEPLOYMENT PLAN

The deployment plan outlines the strategic approach for implementing the project's objectives related to the deployment of the intelligent ecosystem for rare cancer research. It details the tasks, milestones, and activities involved in deploying the capsule infrastructure at the CoEs participating in the project.

The deployment plan encompasses various key components such as data model understanding, test deployments, ETL engines, structured data collection, Federated Learning algorithm integration, NLP data collection, and more. Each task is designed to ensure the successful integration, processing, and utilization of diverse datasets from different CoEs within a federated environment.

Furthermore, the deployment plan emphasizes the importance of meticulous preparation, comprehensive testing, and documentation to establish a secure and stable production environment. It also highlights the significance of post-deployment monitoring and support to address any issues promptly, ensuring a fully operational ecosystem ready for end-users.

Overall, Chapter 3 of the deployment plan provides a detailed roadmap for the phased implementation of the intelligent ecosystem, focusing on data management, integration of advanced technologies, and collaboration among CoEs to advance rare cancer research.

It is like a sub-project in the project aimed to correctly create and run the capsule for the researches outlined in WP8.

3.1 Phase 1: Proof of concept and real testing [Sep23, Aug24]

Phase 1 of the deployment plan, "Proof of concept and real testing," marks the initial step in implementing the intelligent ecosystem for rare cancer research. This phase focuses on foundational tasks (reported in figure 1) to validate the feasibility and functionality of the capsule infrastructure. It begins with CoEs understanding the IDEA4RC data model and mapping their data warehouses to ensure compatibility. Test deployments are conducted to assess the infrastructure's performance within CoEs' environments, with demo data collection showcasing the system's capabilities. ETL processes are tested for seamless data transfer, and structured data is collected from various sources across CoEs. The production capsule deployment involves final validation and testing, data synchronization, and careful deployment scheduling. Integration of Federated Learning algorithms, NLP algorithm definition and

training, governance frameworks, and metadata services are key components of Phase 1. The phase sets a strong foundation for subsequent phases by establishing robust data management practices and testing advanced technologies within the ecosystem for rare cancer research.

Phase 1: Proof of concept and real testing																
Activities				2023/24												
	Start	End	Responsible	1 Sep	2 Oct	3 Nov	4 Dic	5 Jan	6 Feb	7 Mar	8 Apr	9 May	19 Jun	11 Jul	12 Agu
Understanding data model	1	2	UDEU												
Test deployment	2	3	UPM												
Demo data collection	3	5	ENG												
Demo capsule deployment	3	5	UPM												
ETL test	4	9	ENG												
Collecting structured data	5	12	ENG												
Production capsule deployment	6	9	UPM												
ETL on structured data	5	9	ENG												
Federated Learning integration	4	12	IKNL												
NLP alg definition	1	6	FBK/UDEU												
NLP data collection	5	9	UDEU/CLN												
NLP training	5	11	FBK/UDEU												
NLP integration	10	12	UPM/UDEU												
Metadata services	5	12	UDEU												
Governance	8	12	CERTH/TNO												
Definition & design - augmented analytics and multimodal navigator	3	12	UPM/CoEs												
Data mapping at CoEs	4	12	CoEs												
Data quality	4	12	UDEU/IKNL												
Data annotation tool integration	4	7	CLN												
MILESTONES																
MS1 Capsule version 1 up and running in the CoEs with no-personal data																MS1

Figure 1 - Gantt phase 1

3.1.1 Understanding the data model [Sept23, Oct23]

This task focuses on ensuring that CoEs understand the structure of the IDEA4RC data model. CoEs are expected to familiarize themselves with how their data is organized within their data warehouses and compare it to the structure of the IDEA4RC data model. By evaluating the similarities and differences between their existing data structures and the IDEA4RC model, CoEs can identify areas that may require mapping or adjustments for seamless integration.

The desired outcome for CoEs is twofold: first, to define the preferred Point of injection of the Data at CoE, which involves determining the specific entry point or location within their data infrastructure where data will be extracted and integrated into the IDEA4RC ecosystem. Second, CoEs should gain clarity on where to access the relevant data within their Data Warehouses by establishing a mapping process from their Data Warehouse structure to the IDEA4RC data model. This mapping exercise is crucial for ensuring that data can be effectively extracted, transformed, and loaded into the IDEA4RC system, enabling smooth data integration and analysis across the project.

3.1.2 Test deployment [Oct23, Nov23]

The task involves providing CoEs with a test implementation of federated algorithms for evaluating the local deployments of the IDEA4RC system. This demonstration gradually

increases in complexity to align with the Zero Trust Architecture of the capsules, emphasizing security and access control measures.

CoEs are required to test this demo project within their available infrastructure designated for the project by installing and running the tools developed within the projects. In cases where the necessary infrastructure is not accessible, provision will be made for a Virtual Machine (VM) to facilitate the deployment of the capsule for testing purposes.

The expected outcome from CoEs participating in this task is to provide feedback based on their testing experience with the demo deployment. This feedback should include insights on the functionality of the system, any potential issues encountered during the testing process (such as blockages at specific stages), and overall observations regarding the performance and usability of the deployed system. This feedback is crucial for refining and optimizing the deployment process and ensuring that the IDEA4RC system meets the requirements and expectations of the CoEs for effective utilization in rare cancer research initiatives.

3.1.3 Demo data collection [Nov23, Jan24]

The task involves the collection of sample data that aligns with the IDEA4RC data model and the specific point of injection determined by the CoE. During this task, CoEs are required to generate synthetic or fake data that accurately represents the structure and content of their actual data sources, ensuring compatibility with the designated point of injection (PoI) identified in the WP3. The expected PoIs defined in WP3 are:

1. Not standardized data. Within this point of injection CoE needs to provide data in a raw csv format. This point of injection is designed for those CoEs that have no standardized format at sources.
2. IDEA4RC data model. This point of injection is aligned with the ER model of IDEA4RC, data needs to be provided as SQL dump format. This point of injection is designed for those CoEs that decide to adopt the IDEA4RC data model as sources.
3. This point of injection is aligned with the OMOP standard, data needs to be provided as SQL dump format or CSV file for each OMOP clinical table. This point of injection is designed for those CoEs that use OMOP as sources it needs to align to the IDEA4RC data OMOP mapping.

4. This point of injection is aligned with the FHIR standard, data will be loaded using FHIR apis. This point of injection is designed for those CoEs that use FHIR as sources it needs to be aligned to the IDEA4RC FHIR implementation guide defined in the project.

CoEs are expected to provide one or two patient datasets that encompass all the variables and information outlined in the IDEA4RC data model. These datasets should serve as representative samples of the CoE's data, reflecting the diversity and complexity of information that will be integrated into the IDEA4RC ecosystem. By collecting such comprehensive and representative data samples, CoEs can facilitate the testing and validation of data integration processes, ensuring that the system can effectively handle and process real-world data from diverse sources within the CoE's environment.

3.1.4 Demo capsule deployment [Nov24, Jan25]

The task involves the distribution of a demonstration version of the IDEA4RC capsule to CoEs for the purpose of evaluating its local deployment. This demo capsule includes the complete set of Data Services, excluding the Extract, Transform, Load (ETL) engines, and is designed to progressively increase in complexity to align with the Zero Trust Architecture principles of the capsules, emphasizing security and data protection.

CoEs are required to conduct testing of this demo project within their existing infrastructure allocated for the IDEA4RC project to evaluate the feasibility of the deployment in their local environments and solve eventual administrative issue that could the delay the technical development. In cases where CoEs do not have the necessary infrastructure available, provision will be made for a VM to facilitate the deployment and testing of the capsule.

The expected outcome from CoEs participating in this task is to provide feedback based on their testing experience with the demo deployment of the capsule. This feedback should encompass observations on the functionality of the system, any encountered issues or blockages during the testing process (such as at specific stages), and overall assessments of the performance and usability of the deployed capsule. By gathering feedback from CoEs, the project team can identify areas for improvement, address any challenges, and enhance the deployment process to ensure the successful integration and operation of the IDEA4RC capsule within the CoEs' environments.

3.1.5 ETL test [Dec24, Jun25]

In this phase of the project, the primary focus will be on testing the ETL engines at each designated point of injection using the synthetic demo data generated during the Demo data collection phase. This rigorous testing process is essential to ensure that the ETL engines perform effectively across all points of data entry, validating their functionality and reliability under simulated conditions.

The project acknowledges that input from the CoEs may be necessary to refine the ETL processes. As these engines are tested with the demo data, any feedback from the CoEs will be instrumental in identifying areas that require adjustments or enhancements. This could involve pinpointing specific customizations needed to tailor the ETL processes to unique data environments or highlighting the necessity for additional data quality rules to improve the integrity and accuracy of the data.

The expected outcome from the CoEs in this testing phase is a detailed feedback report regarding the requirements of the ETL processes at the various points of injection. This feedback should encompass insights on any specific customizations that may be necessary to optimize the ETL engines for their particular data environments. Additionally, it should address whether further data quality rules need to be incorporated to ensure the data meets the desired standards of accuracy and consistency.

3.1.6 Collecting structured data [Jan24, Aug24]

This task focuses on the advanced collection of structured data, scheduled to occur between January 2024 and August 2024. The primary aim is to gather a sophisticated dataset that aligns with the IDEA4RC data model and the specific points of injection previously established by the CoEs. In this task, the CoEs are required to contribute additional data that accurately represents their source data. Whenever possible, this data should be anonymized to ensure privacy and compliance with data protection standards.

The data collection effort must agree with the points of injection identified in the earlier task, where the CoEs gained an understanding of the data model. This alignment ensures that the collected data is not only comprehensive but also relevant to the various aspects of the IDEA4RC framework. By doing so, the CoEs will be able to provide a dataset that is extensive enough to cover the entire IDEA4RC data model.

The expected outcome for the CoEs in this task is to successfully collect a significant and representative amount of data. This data should thoroughly cover the IDEA4RC data model, thereby ensuring that the dataset is comprehensive and reflective of the CoEs' data environments. The collection of this data is crucial as it will serve as a foundation for further analysis, model validation, and refinement within the IDEA4RC project, ultimately contributing to the project's overall success.

3.1.7 Production capsule deployment [Jan24, Jun24]

This task focuses on deploying the IDEA4RC capsules into a close to production-ready environment (that means a production environment without real data). This crucial phase of the project involves preparing the capsules for real-world application, ensuring they meet all necessary standards and requirements for operational use. The primary responsibility in this task is to transition the capsules from a developmental or testing state to a fully functional, production-ready state. The CoEs are expected to play a key role in this process, meticulously working to implement the deployment within their respective infrastructures. This involves configuring the capsules to operate efficiently and reliably, addressing any potential issues that may arise, and ensuring that all aspects of the deployment are optimized for performance and security. The ultimate goal for the CoEs is to successfully deploy a production-ready version of the capsules, signifying that they are fully prepared for widespread use and capable of performing their intended functions in a live environment. This successful deployment will mark a significant milestone in the project, demonstrating the readiness and robustness of the IDEA4RC capsules for practical application.

3.1.8 ETL on structured data [Jan24, Jun24]

The task is dedicated to advancing the testing of ETL processes integrated into the IDEA4RC capsules. Building on previous ETL engine tests, this task involves utilizing more sophisticated tools to ensure the seamless integration of ETL processes within the capsules. Although a specific timeline is set for this task, it is essential to recognize it as an ongoing activity that must adapt to updates in the data model outlined in Work Package 5 (WP5). This adaptability is crucial as the ETL processes need to align with any changes or enhancements in the data model, ensuring they remain effective and relevant.

The CoEs are expected to execute the ETL engines on the production version of the capsules. This involves running comprehensive tests to verify that the ETL processes are functioning correctly and efficiently within the production environment. The CoEs will need to ensure that the data extraction, transformation, and loading operations integrate seamlessly with the capsules, supporting the overall data flow and maintaining data integrity. The successful execution of this task will demonstrate the capability of the ETL engines to operate in a live setting, providing a robust mechanism for data management and processing in alignment with the evolving data model. This ongoing effort will contribute significantly to the project's aim of delivering a reliable and adaptable data processing framework.

3.1.9 Federated learning integration [Dec23, Aug24]

This task focuses on integrating the Vantage6 engine into the IDEA4RC capsules at each CoE. This task is crucial as it aims to incorporate federated learning capabilities into the capsules, enhancing their ability to perform secure and efficient data analysis across decentralized data sources. The CoEs are responsible for deploying and running the integrated Vantage6 engine within their environments. This process involves setting up the engine, ensuring it functions correctly with the existing systems, and running it to facilitate federated learning operations. During this task, the CoEs are expected to meticulously evaluate the integration, identifying any potential issues or areas requiring customization to better fit their specific needs. Their feedback will be vital in refining the deployment, ensuring that the Vantage6 engine operates optimally within the IDEA4RC framework. The expected outcome from the CoEs is the successful deployment and operation of the integrated Vantage6 engine, accompanied by detailed feedback on the integration process. This feedback will inform necessary adjustments and enhancements, ultimately leading to a more robust and customized federated learning solution within the IDEA4RC capsules. This task not only enhances the technical capabilities of the capsules but also ensures that they are tailored to meet the unique requirements of each CoE, thereby contributing to the overall success and adaptability of the project.

3.1.10 NLP algorithm definition [Sep23, Feb24]

Task 3.1.10, scheduled from September 2023 to February 2024, focuses on defining the NLP algorithm and establishing the format for data collection necessary for NLP training at the CoEs. This task involves leveraging the existing data model and selected terminology to guide

the CoEs in specifying the requirements for developing common NLP algorithms aimed at data extraction. The CoEs are expected to provide detailed requirements that will inform the definition of these algorithms, ensuring they are tailored to meet the specific needs of the data extraction processes within the IDEA4RC framework.

However, it is important to note that this task may not be applicable to every CoE. There may be instances where NLP engines cannot be utilized or are not required due to various constraints or specific operational contexts. In such cases, the limitations or prohibitions regarding the use of NLP need to be thoroughly addressed and justified. This involves providing clear explanations for why NLP cannot be implemented or is unnecessary, ensuring that all decisions are well-documented and reasoned.

The expected outcome from the CoEs is a clear specification of their needs concerning NLP. They must indicate whether they require NLP capabilities and, if so, detail the specific purposes for which NLP will be used. Examples might include detecting dates, identifying entities, or other relevant NLP applications. This specification will guide the development and customization of the NLP algorithms, ensuring they are aligned with the actual needs and capabilities of each CoE. This task is essential for creating a robust and effective NLP framework within the IDEA4RC project, tailored to the diverse requirements of the participating CoE.

3.1.11 NLP training [Jan24. Aug24]

This task is dedicated to the training of NLP algorithms. However, it is acknowledged that not every CoE may execute this task, as some may not require or be able to use NLP engines for various reasons. At the project level, NLP engines are expected to be trained independently of the CoEs, ensuring a standardized approach across the project.

In exceptional cases, extreme situations may necessitate that NLP training be conducted directly by a CoE. If this occurs, the CoE is responsible for executing the entire NLP training process within their own infrastructure. The expected outcome from the CoEs regarding this task is twofold. Firstly, if a CoE is utilizing NLP, they need to determine whether they will externalize the data to either FBK or UDEU for training or if they will conduct the training in-house. This decision is crucial as it impacts the workflow and resource allocation for NLP training.

If the CoEs choose to externalize the training to FBK or UDEU, they must share the necessary data in an anonymized and annotated form. This shared data is essential for the external entities

to effectively train the NLP engines, ensuring that the algorithms are accurately tuned to the specific data requirements of the CoEs. This collaborative approach aims to leverage centralized expertise while ensuring that the data privacy and contextual relevance are maintained. Overall, the successful training of NLP engines, whether done in-house or externally, is critical for enhancing the data processing capabilities within the IDEA4RC project, enabling more efficient and accurate data extraction and analysis.

3.1.12 Metadata services [Jan24, Aug24]

The task focuses on establishing a standardized method for describing data within the intelligent ecosystem project. This involves creating a framework of services that will detail information about datasets, quality attributes, and data labels present at CoEs. Additionally, the task aims to ensure alignment with the EHDS guidelines, as outlined by the ehds2pilot project, to facilitate integration with the HealthData@EU infrastructure for the secondary use of data within the EHDS context.

The expected outcome for CoE participating in this task is to successfully implement and deploy the metadata services within their respective capsules. By deploying these metadata services, CoEs will enhance the organization and accessibility of data, ultimately contributing to the project's overarching goal of improving data management and utilization for rare cancer research.

3.1.13 NLP data collection [Jan24, Jun24]

The task pertains to the collection of textual data from CoEs participating in the intelligent ecosystem project. It is important to note that this task may not be carried out by every CoE, especially if there are constraints preventing the use of NLP engines or if such capabilities are deemed unnecessary for specific operational contexts.

For CoEs that will utilize NLP algorithms, the task requires them to provide annotated data in alignment with the format required for training the NLP algorithms. This data collection process is essential for customizing the NLP algorithms to suit the language and specific needs of each CoE. The expected outcome for CoEs involved in this task includes identifying and extracting free text from their Data Warehouses (DWH) based on the predefined IDEA4RC Data Model variables. Furthermore, CoEs are tasked with anonymizing the extracted free text and

generating a set of annotated data that adheres to the Annotation Guidelines, facilitating the language customization of NLP algorithms.

3.1.14 Governance [May24, Aug24]

The task focuses on establishing a governance framework within the intelligent ecosystem project to ensure legal access to data through a structured process involving data permits. The primary objective of this task is to define and implement a governance mechanism that regulates the access to data in compliance with legal requirements and data protection standards.

The expected outcome for CoEs participating in this task includes providing input on the legal requirements specific to their organization. Additionally, CoEs are expected to deploy a data permit service and dashboard, which will serve as essential tools for managing and monitoring data access permissions within the project. By fulfilling these outcomes, CoEs will contribute to the establishment of a robust governance structure that safeguards data integrity and privacy while facilitating authorized access for research and analysis purposes.

3.1.15 Definition and design of the analytics and multimodal navigators [Nov23, Aug24]

The task is focused on the development and validation of the analytical tools within the intelligent ecosystem project. The primary objective of this task is to design and validate the functionalities of the augmented analytics and multimodal navigators to support the use cases defined in WP8.

The expected outcome for CoEs involved in this task includes confirming the stakeholders engaged in each step of the user journey. CoEs are also required to define the needs and expected functionalities of the navigator, ensuring alignment with project requirements. Furthermore, CoEs are expected to validate the navigator requirements and collaborate in the co-design and validation of sketches for all functionalities of the navigator. By achieving these outcomes, CoEs will contribute to the development of user-friendly and effective analytical tools that support data exploration and decision-making within the intelligent ecosystem for WP8 use cases.

3.1.16 Data mapping at CoEs [Dec23, Aug24]

The task is focused on the process of aligning local data at CoEs with the established IDEA4RC data model. The primary objective of this task is to map the existing data within each CoE to ensure compatibility and integration with the standardized IDEA4RC data model.

The expected outcome for CoEs participating in this task includes the collection of data from one or two patients that encompass the entire IDEA4RC data model, providing a representative sample of the CoE's data landscape. CoEs are also required to identify critical variables that need to be mapped onto the IDEA4RC data model to facilitate seamless data integration. Additionally, CoEs are expected to collect their entire local dataset in accordance with the IDEA4RC data model and the designated point of injection, ensuring consistency and alignment with project requirements. By achieving these outcomes, CoEs will contribute to the harmonization of data across the project, enabling effective data processing and analysis within the intelligent ecosystem.

3.1.17 Data quality [Dec23, Aug24]

The task is focused on establishing and upholding processes and standards to ensure the high quality of data throughout its lifecycle within the intelligent ecosystem project. This task encompasses maintaining data quality at the source level, within CoEs, at aggregated levels, and within cohorts. It is crucial to align these efforts with the insights and recommendations provided by the QUANTUM project to harmonize data quality and utility labels of CoEs with the EHDS requirements.

The activities involved in this task include data validation and evaluation based on the quality framework outlined in Work Package 4. The overarching goal is to provide reliable and trustworthy data to support informed decision-making processes and to meet the necessary compliance standards. The expected outcomes of this task revolve around defining a robust data quality framework, which will require input and expertise from CoE specialists. Furthermore, the validation of this framework at the CoEs level is essential to ensure that data quality standards are consistently met across all project stakeholders. By achieving these outcomes, the project aims to enhance data integrity, reliability, and usability for research and analysis purposes within the intelligent ecosystem.

3.1.18 Data annotation tool integration [Dec23, Apr24]

The task focuses on the integration of a data annotation tool within the intelligent ecosystem project. The primary objective of this task is to seamlessly incorporate the Clininote tool for data annotation (https://clininote.com/clininote-system/), ensuring alignment with the IDEA4RC data model.

The expected outcome for CoEs participating in this task is the successful integration of the Clininote tool in accordance with the established IDEA4RC data model. It is essential for the integration process to be updated in alignment with any new versions of the data model that may be introduced from Work Package 5. By achieving this outcome, CoEs will enhance their data annotation capabilities and ensure that the tool is effectively integrated into the project's data management framework, facilitating efficient data annotation and analysis processes within the intelligent ecosystem.

3.1.19 Milestone 1: Capsule version 1 up and running in the CoEs with no-personal data

Milestone 1 of the pilot deployment plan, "Capsule version 1 up and running in the CoEs with no-personal data," is the current milestone that is in delivery at time of writing the deliverable, the primary objective of the milestone is to establish the foundational infrastructure of the intelligent ecosystem for rare cancer research. This milestone serves as the initial step in implementing the project's vision of enhancing data governance, sharing, and federated analysis within the CoEs.

The significance of Milestone 1 lies in the successful deployment and activation of the first version of the capsule within the CoEs, ensuring that the system is operational and accessible to stakeholders. Importantly, the decision to start with no-personal data underscores the project's commitment to data privacy and security from the outset. By utilizing non-personal data during this phase, the project can validate the functionality of the infrastructure while safeguarding sensitive patient information.

By achieving Milestone 1, the project sets the stage for subsequent phases by providing the CoEs with a platform to familiarize themselves with the capsule environment and its basic features. This early deployment allows for testing the system's performance, scalability, and usability, enabling stakeholders to provide feedback for iterative improvements.

Furthermore, Milestone 1 acts as a critical building block for the integration of real-world data in later stages of the project. It establishes a solid foundation for the CoEs to understand the

infrastructure's capabilities and prepares them for the complexities of handling sensitive healthcare data in a secure and compliant manner.

Overall, Milestone 1 represents the initial milestone in a series of planned developments aimed at creating an intelligent ecosystem that supports advanced analytics, data sharing, and collaborative research in the field of rare cancers. It signifies the project's commitment to responsible data management practices and sets the trajectory for the successful implementation of subsequent milestones.

3.2 Phase 2: Implementation and scaling [Sep24, Sep25]

Phase 2 of the pilot deployment plan, known as "Implementation and scaling," represents the final step in the development of the intelligent ecosystem for rare cancer research. This phase builds upon the foundational work of Phase 1 and aims to expand the capabilities of the ecosystem within a federated environment with real data.

In this phase, key tasks include initiating the deployment of a fake data capsule to simulate system performance, preparing the production environment for operational deployment, and deploying the production capsule with real data to validate its readiness for live usage. Additionally, there is a focus on testing and optimizing the ETL engines for efficient data processing, gathering structured data from various sources in the CoEs, and integrating advanced Federated Learning algorithms to enhance collaborative research capabilities.

Furthermore, tasks involve collecting NLP data for text analysis, training NLP models to extract meaningful insights, and integrating NLP capabilities into the ecosystem for enhanced data analysis. Metadata services are implemented to manage data effectively, governance frameworks are integrated to ensure compliance and data security, and advanced analytics and navigation tools are incorporated for improved data exploration and visualization.

The phase also includes continuing activities started in phase 1 such as data mapping at CoEs to facilitate seamless integration of data and tools, ensuring data quality standards are met for reliable research outcomes, and integrating data annotation tools for enhanced data understanding and analysis. Overall, Phase 2 aims to advance the capabilities of the intelligent ecosystem, promote collaboration among CoEs, and scale the project for broader impact in rare cancer research. The tasks (reported in figure 2) outlined in this phase are designed to enhance data management, integrate advanced technologies, and foster collaboration to drive meaningful insights and advancements in the field of rare cancer research. These expected

tasks may change depending on the new detected risks and requirements that may not be considered at time of writing the deliverable. The changes that eventually may occur will be reported in D6.2.

Phase 2: Implementation and scaling				2024/25												
Activties	Start	End	Responsible	1 Sep	2 Oct	3 Nov	4 Dic	5 Jan	6 Feb	7 Mar	8 April	9 May	#####	11 Jul	12 Agu	13 Sept
Fake data capsule deployment	1	2	UPM/ENG/IKNL													
Preparation of production environment	3	5	ENG													
Production capsule deployment	3	5	UPM													
ETL engines	4	9	ENG/UPM													
Collecting structured data support	5	12	ENG													
Novel Federated Learning algorithm integration	4	12	IKNL													
NLP data collection	5	9	UDEU/CLN													
NLP training	5	11	FBK/UDEU													
NLP integration	10	12	UPM/UDEU													
Metadata services integration	5	12	UDEU													
Governance integration	8	12	CERTH/TNO													
Integration of augmented analytics and multimodal navigator	3	12	UPM/CoEs													
Data mapping at CoEs	4	12	CoEs													
Data quality services integration	4	12	UDEU/IKNL													
Data annotation tool support	4	7	CLN													
MILESTONES																
MS2 Capsule version 2 up and running in the CoEs with real data										MS2						
MS2 Capsule version 3 up and running in the CoEs with real data, VA an governance integrated																MS3

Figure 2 - Gantt phase 2

3.2.1 Fake data capsule deployment [Sept24, Oct24]

This task involves the creation and deployment of a fake data capsule, which is essentially a collection of synthetic data used to mimic real-world data interactions. The purpose is to test and validate system functionalities, security measures, and data handling processes without the risks associated with using actual sensitive data.

3.2.2 Preparation of production environment [Sept24, Feb25]

The preparation of the production environment involves setting up the necessary infrastructure, installing required software, and configuring system settings to ensure optimal performance and security. This process includes provisioning servers and databases, migrating and setting up data, and conducting comprehensive testing to validate the setup, all elements and components will be described in deliverables associated to WP3 and WP4. Additionally, thorough documentation and the establishment of a robust backup and recovery plan are essential to maintain stability and ensure business continuity. This meticulous preparation ensures the production environment is secure, stable, and ready for live deployment of applications.

3.2.3 Production capsule deployment [Feb25, Mar25; Jul25, Sep25]

The production capsule deployment involves final validation and testing to ensure system functionality. The deployment is carefully executed according to a predefined schedule

following a continuous integration and continuous delivery process with roll backs to minimize disruptions. In post-deployment, the system is closely monitored to identify and resolve any issues promptly, ensuring a stable and fully operational live environment ready for end-users. Procedures and activities for CoE to perform assisted deployments will be established.

3.2.4 ETL engines [Sep24, Feb 25; Apr25, Aug25]

This task follows what started during the previous ETL phase. It will provide exhaustive test and support for the foundational data management tasks necessary for successful deployment and operation of the research platform, ensuring that data from various sources across different CoEs can be integrated, processed, and utilized effectively for rare cancer research. This task follows from phase 1.

3.2.5 Collecting structured data [Sep24, Dec24]

The primary goal of this task is to systematically gather structured data from various sources across the participating CoEs involved in the IDEA4RC project. Structured data refers to data that is organized in a predefined format, often using standardized codes, classifications, or schemas. This task follows from phase 1.

3.2.6 Novel Federated Learning algorithm integration [Sep24, Jul25]

The primary goal of this task is to integrate novel Federated Learning (FL) algorithms developed in WP4 by IKNL into the IDEA4RC project's infrastructure to support the use cases of WP8. Such federated Learning algorithms could include machine learning techniques enabling training models across decentralized edge devices (in this case, different CoEs) without centrally aggregating the data of statistical algorithms executed in distributed environments. This approaches preserve data privacy while allowing collaborative model training. This task follows from phase 1.

3.2.7 NLP data collection [Sept24, Dec24]

The main objective of this task is to gather textual data from various sources across the CoEs participating in the IDEA4RC project. This textual data is essential for training and refining NLP models. NLP techniques will be used to extract valuable insights and structured information from unstructured text data related to rare cancers. This task follows from phase 1.

3.2.8 NLP training [Oct24, Jan25]

The primary objective of this task is to train NLP models using the collected textual data from various sources across CoEs. These models are designed to automatically extract structured information, identify patterns, and derive meaningful insights from unstructured text data related to rare cancers. This task follows from phase 1.

3.2.9 NLP integration [Nov24, Jan25; Apr25, Jul25]

This task aims to seamlessly integrate the NLP models and tools developed during the NLP training phase into the operational environment of CoEs within the IDEA4RC project. This integration aims to enhance the project's ability to process and extract semantics from unstructured textual data related to rare cancers. This task follows from phase 1.

3.2.10 Metadata services [Sept24, Jul25]

The objective of this task is to develop and deploy metadata services that facilitate the organization, discovery, and management of data assets provided by CoEs within the IDEA4RC project. Metadata services play a crucial role in enhancing data interoperability, accessibility, and usability and alignment of CoE data space with EHDS discoverability requirements. This task follows from phase 1.

3.2.11 Governance integration [Sept24, Jul25]

This task will establish and integrate governance mechanisms that ensure ethical, legal, and regulatory compliance in the management and sharing of healthcare data across multiple CoEs participating in the IDEA4RC project developed in WP7.

3.2.12 Integration of augmented analytics and multimodal navigator [Sep24, Jul25]

The integration of augmented analytics and multimodal navigator developed in WP6 aligns with IDEA4RC's overarching goal of leveraging advanced technologies to accelerate rare cancer research, improve clinical decision-making, and enhance patient care outcomes as expected in WP8. By providing researchers with powerful tools for data exploration and analysis, the project aims to foster innovation and collaboration in the field of oncology.

3.2.13 Data mapping at CoEs [Sept24, Jan25]

The objective of this task is to establish a standardized approach for mapping and harmonizing data from various sources across multiple CoEs participating in the IDEA4RC project. This includes extracting data at original sources, evaluate the associated quality indicators and semantic mappings in agreement with the PoIs defined in the project to create the data space of the CoE IDEA4RC capsule. This task follows from phase 1.

3.2.14 Data quality [Sep24, Jan25]

This task will follow the same task of the previous phase, it involves implementing and maintaining processes and standards to guarantee high data quality throughout the data lifecycle (at sources, at CoE, at aggregated CoEs and at cohort levels). Activities include data validation and evaluation against the quality framework defined in WP4.

Data annotation tool integration [Sept24, Oct24]

The objective of this task is to deploy and integrate data annotation tools at each participating CoEs within the IDEA4RC project that are using the NER algorithms developed in WP5. These tools facilitate the systematic labelling and tagging of structured and unstructured data to enhance data quality, interoperability, and usability for research and clinical applications and later train of the NER algorithms. This task follows from phase 1.

3.2.15 Milestone 2: Capsule version 2 up and running in the CoEs with real data

Milestone 2 in marks a significant advancement in the project's journey towards enhancing rare cancer research through advanced technological deployment. This milestone is pivotal as it transitions from the initial phases of testing and prototyping to the full-scale implementation of infrastructure across multiple CoEs.

The primary objective of Milestone 2 is to deploy the second version of the capsule infrastructure at each CoE involved in the project. This infrastructure includes a sophisticated framework designed to handle diverse datasets. The deployment process entails configuring and integrating updated software components and tools necessary to support data processing, analysis, and collaboration within a federated environment.

Key to this milestone is the integration of real-world data into the capsules. This involves populating the deployed infrastructure with actual datasets sourced from the CoEs. These

datasets are meticulously standardized to adhere to established data models, vocabularies, and ontologies (such as the IDEA4RC data model, OMOP, and FHIR), ensuring consistency and interoperability across different CoEs. By standardizing data, the project enables federated analysis while upholding stringent data privacy and security protocols.

Testing and validation are crucial components of Milestone 2. Rigorous testing procedures are conducted to verify the functionality, performance, and scalability of the deployed infrastructure. This includes testing data ingestion capabilities, processing workflows, query mechanisms, and data security measures to ensure robustness and reliability in handling sensitive healthcare data. User acceptance testing (UAT) engages stakeholders, including researchers and clinicians, to validate the usability and effectiveness of the capsule environment, gathering feedback to enhance user interfaces, governance framework and analytical tools.

Furthermore, governance and compliance frameworks are integrated into the capsule infrastructure to uphold ethical guidelines, data protection regulations (e.g., GDPR), and institutional policies across CoEs. Mechanisms for data access control, privacy-preserving data sharing, and accountability in data management practices are implemented to safeguard patient confidentiality and trust.

Documentation plays a critical role in supporting the operationalization of capsule version 2. Comprehensive documentation is created detailing the setup, configuration, and operational procedures necessary for managing and maintaining the infrastructure. Knowledge transfer activities, such as training sessions and workshops designed for CoE staff and stakeholders with the skills needed to effectively utilize the infrastructure for conducting research, clinical trials, and collaborative studies on rare cancers, will be performed.

Achieving Milestone 2 signifies substantial progress towards the establishment of the IDEA4RC secure federated data processing environment that is the main goal of the IDEA4RC project. It establishes a solid foundation for subsequent phases, paving the way for integrating advanced analytics, novel algorithms, and scaling up data-driven initiatives across the federated network of CoEs. Ultimately, Milestone 2 aims to accelerate rare cancer research, foster international collaboration, and improve healthcare outcomes through innovative data-driven approaches.

3.2.16 Milestone 3: Capsule version 3 Virtual Assistant (VA) and governance integrated

Milestone 3 represents the outcome in the project's evolution, marked by the integration of advanced functionalities and governance frameworks within the capsule environment deployed across CoEs. This milestone underscores the project's commitment to enhancing both technological capabilities and ethical standards in rare cancer research.

Central to Milestone 3 is the integration of Virtual Assistant (VA) capabilities within the capsule infrastructure. The VA, powered by artificial intelligence (AI) and NLP, serves as an intelligent interface designed to streamline user interactions and optimize access to complex datasets and analytical tools. By leveraging AI-driven solutions, the VA facilitates seamless navigation, query handling, and real-time support for researchers, clinicians, and stakeholders for performing federated analysis. This integration aims to democratize access to advanced analytics, empowering users to extract meaningful results and accelerate data-driven decision-making processes in rare cancer research.

Simultaneously, Milestone 3 focuses on strengthening governance frameworks within the capsule environment. Governance integration ensures ethical compliance, data protection, and regulatory adherence across all participating CoEs. Such integration includes stringent data privacy protocols, encryption standards, and anonymization techniques to safeguard patient confidentiality and mitigate risks associated with unauthorized access or breaches.

Moreover, governance integration encompasses ethical considerations in rare cancer research, emphasizing informed consent, fair data usage, and adherence to ethical standards. This includes regular monitoring, auditing, and evaluation of compliance with regulatory requirements, ensuring alignment with established guidelines and best practices. Capacity-building initiatives provide training and educational resources to CoE personnel and stakeholders, fostering a culture of responsible data management and promoting awareness of ethical guidelines in healthcare research.

By integrating VA capabilities and robust governance frameworks into the capsule infrastructure, Milestone 3 represents a significant advancement toward enhancing the platform's functionality, usability, and compliance standards. It aims to foster collaborative research efforts among CoEs, facilitate data-driven insights, and ultimately contribute to improved healthcare outcomes for patients affected by rare cancers. This milestone

underscores IDEA4RC's commitment to innovation, ethical integrity, and impactful advancements in rare cancer research.

4 ADAPTABLE DEVOPS CULTURE FOR DEPLOYEMENT, MANAGEMENT, SUPPORT AND EXECUTION

From the operational viewpoint, the main challenge for conducting the deployment plan illustrated in the previous section, is the heterogeneity intrinsic in the IDEA4RC Pilot, that spans 11 different CoEs, belonging to the EURACAN network, located in 8 different member states. This entails the need to appropriately tackle specificities in several directions, e.g.:

- Different technical bases, at the IT department of each CoE
- Different approaches to data management at each CoE
- Different regulatory contexts, depending on the CoE country and Regions (differences in GDPR application in EU MSs, additional local rules, etc.)

This requires an adaptable approach, that allows to deploy the IDEA4RC platform while at the same time flexibly adjusting to such differences:

- *Technically*, this has been implemented by allowing different technical options in the deployment plan tasks, as illustrated in the previous section (e.g. different options for data ingestion points, for NLP application, etc.), that can be selected by CoEs as they suit best.
- *Operationally*, this has been implemented by enacting an appropriate management, support and monitoring framework, which is illustrated in this section.

4.1 Management

Management of Pilot deployment revolves around two elements:

- Identification of roles and contact persons for Pilot deployment at each CoE
- Establishment of communication means

These are presented in detail below.

4.1.1 Roles and contact persons: actors' map and WG on Pilot Deployment

Each CoE has been invited to appoint several *representative roles*, charged to interact with the rest of the IDEA4RC project team, as illustrated in the "Actors' map" represented in Figure 3:

- *Overall CoE representative*, in charge of overall representation of the respective CoE (e.g. on contractual matters, relating to budget, project objectives, data management,

etc.), interacting in particular with IDEA4RC project coordination and management roles

- CoE *representative for business aspects*, in charge of project dissemination and exploitation of results, interacting with the WP11 leadership
- CoE *representative for legal and ethical aspects*, in charge of managing the respective aspects, such as for instance handling the signature of data agreements in compliance with the GDPR and other relevant local regulations, and approval of data governance rules.
- CoE *representative for technical aspects*, in charge of the respective aspects, including data integration, ETL, data mapping, data extraction, NLP, entity mapping, AI and federated data analysis, platform deployment and technical support. They interact with the project technical team (working in the corresponding technical areas).

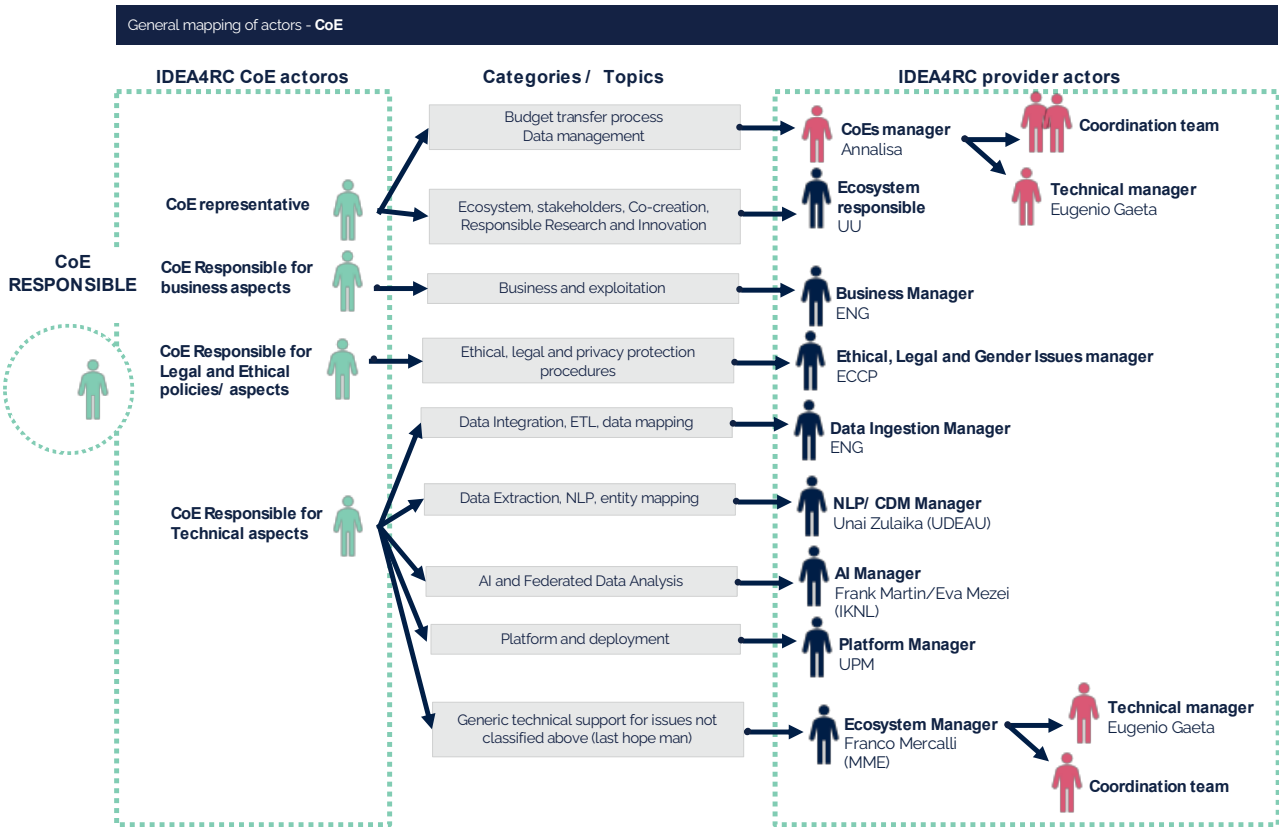

Figure 3 - Actors' map

With specific reference to Pilot deployment planning and management, the last role is the main contact point for CoEs and, as such, they interact with WP9 leadership, which in turn interacts with the project Coordination team and the Technical Management team.

This group of people - CoE representatives for technical aspects, representative of WP9 leadership, representatives of the Coordination and Technical Management teams – plus the representatives of two other WPs – WP6 and WP8 – which have important implications in Pilot deployment, collectively form the "Working Group (WG) on Pilot Deployment", which is the main management body for this topic, and is chaired by WP9 leadership.

Occasionally, when Pilot deployment matters touch the respective remits, other CoE representative roles or specific technical partners might be involved in the WG on Pilot Deployment, on request by the WG chair (for example: when deployment of the Governance Layer is addressed, the CoE representatives for legal and ethical aspects as well as the technical partners participating in WP7, might be involved on request).

4.1.2 Communication

In addition to conventional email and phone communication, the members of the WG on Pilot Deployment, discussed in the previous subsection, interact through two additional instruments:

- *Slack channels* (see Figure 4). Several channels are defined to focus on specific sub-areas of Pilot deployment work. Some of them are dedicated to technical partners, to discuss the way individual platform modules are developed and how they are integrated into the overall IDEA4RC platform for deployment, while others are dedicated to interaction among the technical team and CoE representatives, on matters specifically related to the deployment of the pilot platform at CoEs. These latter communication channels are a crucial element in ensuring the flexibility of Pilot deployment planning, when adapting to the specific needs of specific CoEs.

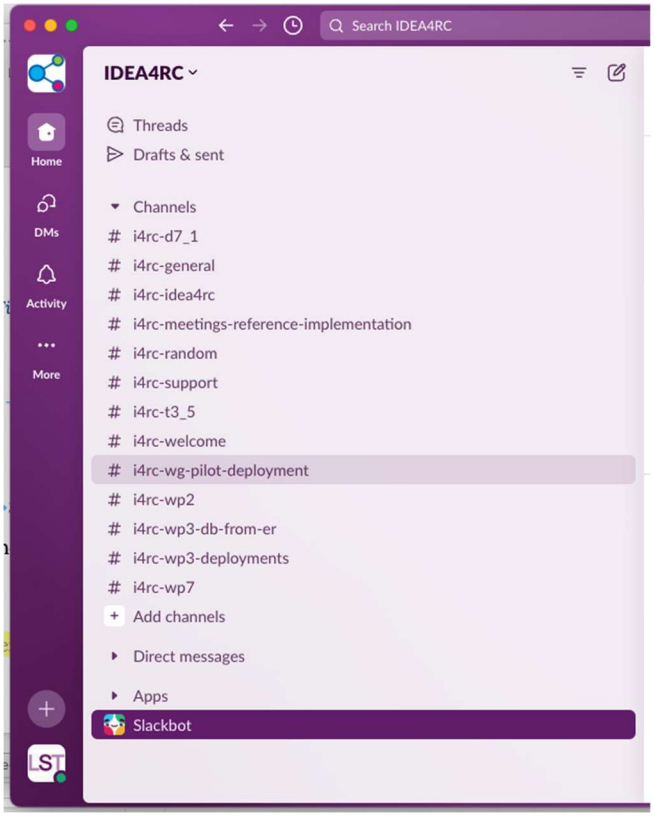

Figure 4 - Slack channels

- A *shared repository* for the WG on Pilot Deployment (see Figure 5), accessible to all members of the WG. The repository contains operational information, such as members contact lists; TODO lists (see later); meetings calendar invites, agendas and materials (see later), and "how to" documents, as well as technical content, regarding specific aspects of the deployment activities. The repository is an essential hub for all WG members – for CoEs and for technical partners alike – to exchange information and knowledge.

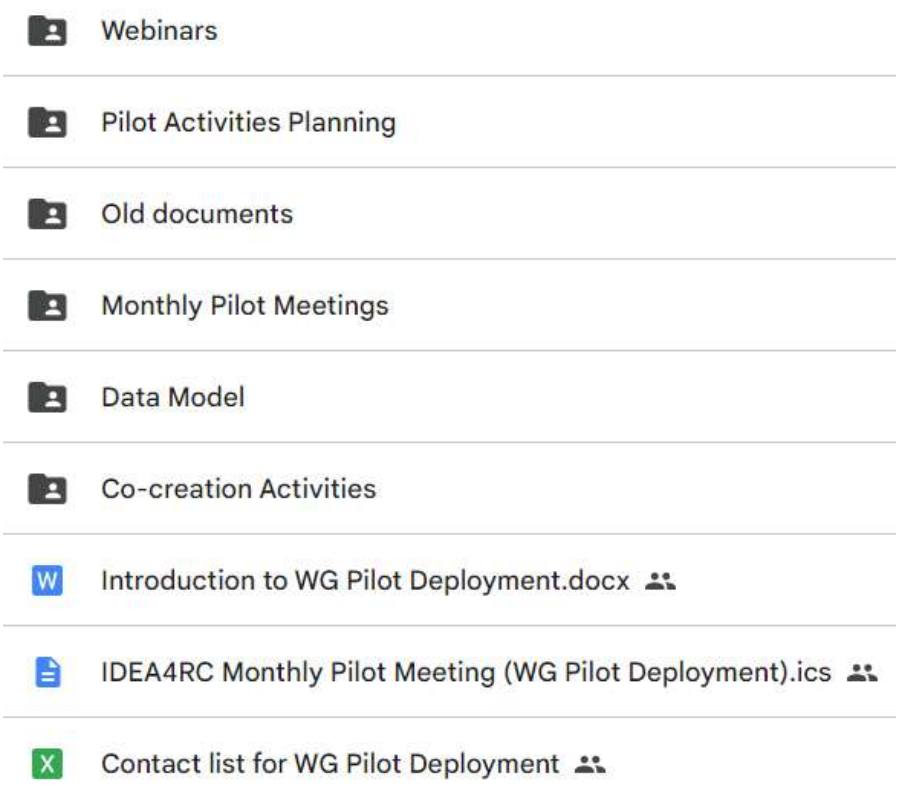

Figure 5 - Shared repository for the WG on Pilot Deployment, example excerpt

4.2 Support and monitoring

In addition to the definition of the involved actors, the institution of a management WG and the establishment of communication tools, as illustrated in the previous subsection, the adaptable Pilot deployment strategy of IDEA4RC is underpinned by *continuous support and monitoring*, taking into considerations the specificities and constraints at each participating CoE, as illustrated below.

4.2.1 Periodic and ad hoc meetings

A periodic *Monthly Pilot Meeting* is convened every last Monday of the month, lasting 1h30 (from 10:00 CET to 11:30 CET) to provide a venue for the WG on Pilot Deployment and periodically conduct the following tasks:

- Technical partners present to CoE relevant next steps on Pilot deployment activities, to be carried out in the next month (this might be supplemented by delivery of specific educational webinars, see later on)

- CoEs may request clarification from technical partners on issues and/or unexpected difficulties that possibly arose in each local context when trying to implement the steps planned for the previous months
- WG leadership conducts a periodic overall check on work advancement and pending issues at each CoE (see next subsection)

For each Monthly Pilot Meeting, a relevant sub-folder is created in the shared repository of the WG on Pilot Deployment, containing the meeting agenda, the attendance list, any slide decks and other materials presented during the meeting, and (when available) video and audio recording of the meeting itself (an example is presented in Figure 6).

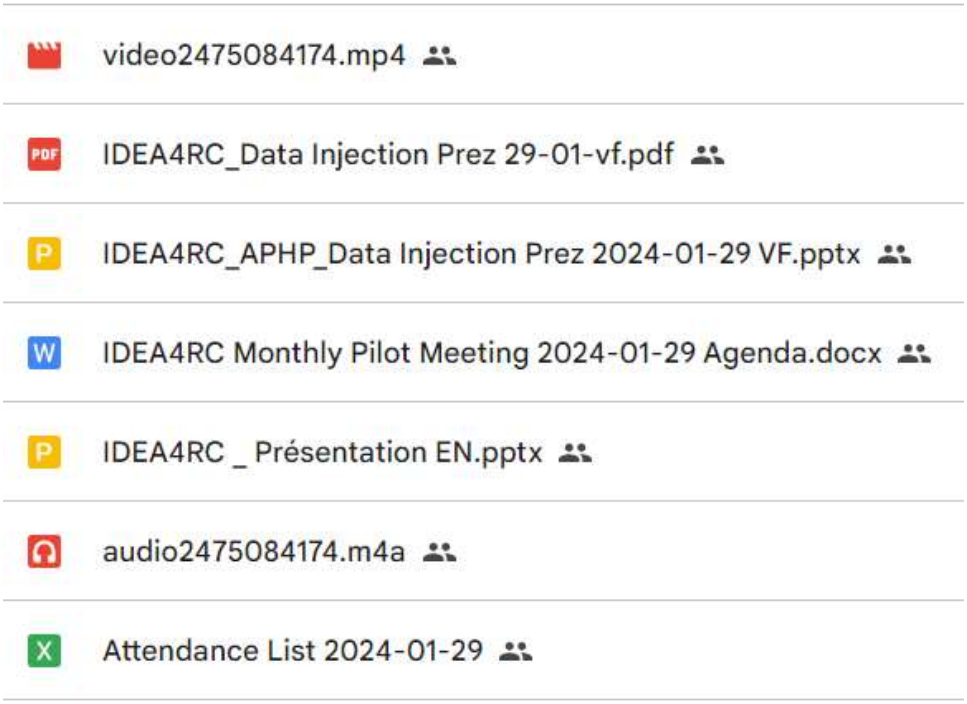

Figure 6 - Example of materials for a Monthly Pilot Meeting

At the time of this writing, the following Monthly Pilot Meetings have been held:

- 1st Meeting, on Oct 23rd, 2023
- (November Meeting not held, as the IDEA4RC 3rd Plenary Meeting was held on Nov 22-23rd 2023)
- 2nd Meeting, on Dec 11th, 2023 (anticipated, due to the upcoming holiday period)
- 3rd Meeting, on Jan 29th, 2024
- 4th Meeting, on Feb 26th, 2024
- 5th Meeting, on March 25th, 2024

- (April Meeting not held, as the IDEA4RC 4th Plenary Meeting was held on Apr 29-30th, 2024)
- 6th Meeting, on May 27th, 2024
- 7th Meeting, on Jun 24th, 2024
- 8th Meeting, on Jul 29th, 2024
- (August Meeting not held, due to the holiday period)

The Monthly Pilot Meetings will continue to be held with the above periodicity until the end of WP9, planned for M45 (May 31st, 2026).

In addition to the periodic Monthly Pilot Meetings, additional meetings might be called by the WG Leadership, in agreement with Coordination and Technical Management teams, in order to address specific issues that need to be tackled. This is another important mechanism to contribute to ensure that specificities at CoEs are managed appropriately, including through 1-to-1 interactions.

4.2.2 Activity monitoring

Advancement on Pilot deployment activities presented in section 3 is periodically discussed in the Monthly Pilot Meetings, as described in the previous subsection, and is continuously monitored through a shared "TODO list" spreadsheet, an excerpt of which is presented in Figure 7 below. The TODO list is accessible to all members of the WG on Pilot Deployment, within the WG's shared repository (illustrated previously).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Call for action (with links)	Expected dead line	Profile of actor	Global Status	INT			CLB			APHP			IIS-FJD	
2					*Notes from Mgmt*	CoE Report	Status	*Notes from Mgmt*	CoE Report	Status	*Notes from Mgmt*	CoE Report	Status	*Notes from Mgmt*	CoE Report
						with the variables of H&N IDEA4RC data model									
4	**NLP CLN Tool**: Report on status of integration of CNL annotation tool	2024-04-29	Technical staff	Open	*INT is interacting with CLN to solve technical issues*	We are waiting for CLN dictionaries to start annotation for sarcoma & H&N	closed	*Not using NLP*	N/A	closed	*If needed, will use own annotation tools*	N/A	closed	*Using own annotation tools*	N/A
5	**NLP for CoEs not using CLN / 1**: REPORT, will you share with FBK the annotations done with your own annotation tool?	2024-05-20	Technical staff and legal staff	Open	*Using CLN*	N/A	closed	*Not using NLP*	N/A	closed	*If needed, will use own annotation tools*	not sure about using NLP. If so, if there is any interest in sharing with FBK, we might do so.	closed	*Using own annotation tools*	

Figure 7 - Excerpt of TODO List for WG on Pilot Deployment monitoring

The spreadsheet is structured as follows:

- There is one row for each current activity of the Pilot Deployment Plan to be monitored. Activities are added as soon as they need to be acted upon by CoEs and are hidden once they have been satisfactorily achieved by all CoEs, in such a way that only the actions CoEs should currently focus on are visible. In particular, each row includes:
    - A description of the activity (full details are provided during the Monthly Pilot Meetings, as discussed previously, when the activity is added anew)
    - An overall deadline for the activity to be completed by CoEs
    - An indication of the type of actor at CoEs that should deal with the activity (see actors' map, previously discussed)
    - A global status, indicating if the activity is still open (i.e. some CoEs still need to complete it) or if it is closed (all CoEs have satisfactorily completed it).
- A column for each CoE, where the specific status of each activity at the relevant CoE, as well as any special issue applying to such CoE, is reported. Each column includes:
    - An overall status flag that can assume one of the following values:
        - "To do" (colour: orange): the activity has been just added by the WG on Pilot Deployment management and the CoE is invited to address it at earliest convenience.
        - "Ongoing" (colour: blue): the activity is acknowledged as being addressed by the CoE, which is currently working on its completion
        - "Pending" (colour: red): the activity is past its deadline and the CoE is expected to report in the spreadsheet what issues stand in the way of completion, so that the whole team can jointly examine the situation and find appropriate solutions
        - "Closed" (colour: green): the activity has been satisfactorily completed by the CoE
    - A text area where the WG on Pilot Deployment management can report specific notes, directed to the respective CoE. This is an important element, as it allows to deal with specificities and particular local issue and/or constraints, that must be addressed individually for the CoE, enacting the necessary adaptability of the Pilot Deployment plan, as previously mentioned

- A text area where the CoE can also report about such specificities and constraints, with the same purpose as the above bullet, enacting a discussion among the WG on Pilot Deployment management and the CoE, conducive to a shared contingency planning process, that ensure effective adaptability

As previously mentioned, the TODO List spreadsheet is jointly checked at least monthly, in the frame of the Monthly Pilot Meetings. However, it is also continuously updated by CoEs and continuously monitored by the WG management, ensuring timely detection of issues. Special 1-to-1 checks with individual CoEs are also possible, in case of urgency or in case of individual cases that must be addressed separately with the respective CoE.

4.2.3 Webinars

In order to enable CoEs to optimally and efficiently start and complete the various Pilot Deployment activities illustrated in section 3, technical partners often need to educate CoEs (especially CoEs representatives for technical aspects) on the various IDEA4RC platform components to be deployed at their premises, how they work jointly with other IDEA4RC components, and how they must be integrated within each CoE local infrastructure.
In order to do this, relevant Webinars are prepared and delivered by technical partners to CoEs as needed (Figure 8).

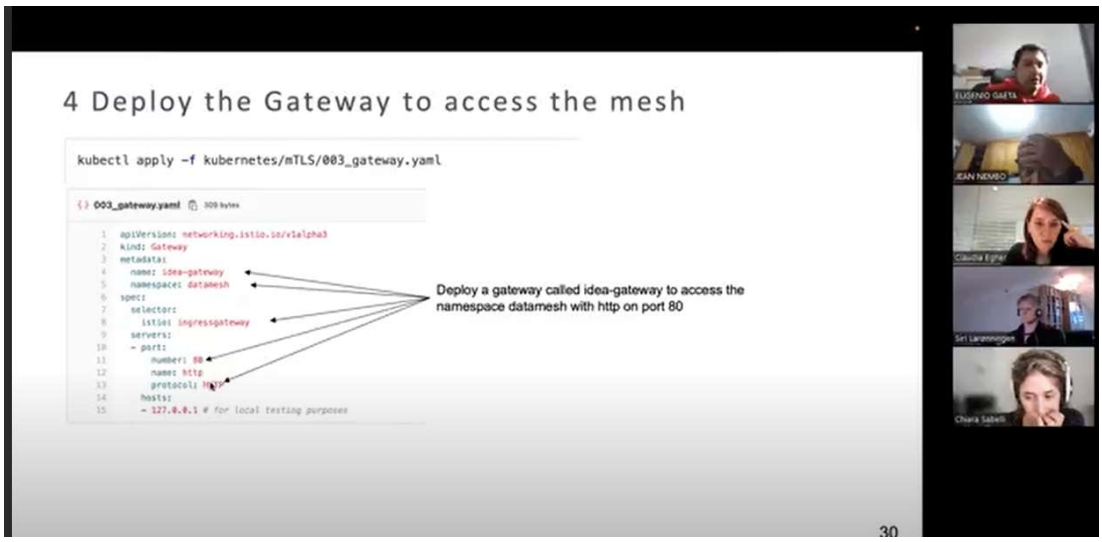

Figure 8 - Screenshot form a Pilot Deployment Webinar

As an example, at the time of this writing, the following 9 Webinar have been organized:

- Nov 6th, 2023: Webinar on Capsule Test Deployment
- Nov 13th, 2023: Webinar on Data Model – HNC & Sarcoma

- Jan 15th, 2023: Webinar on Data Ingestion at CoE / 1
- Jan 22nd, 2024: Webinar on Data Ingestion at CoE / 2
- Feb 5th, 2024: Webinar on Data Ingestion at CoE / 3
- Mar 11th, 2024: Webinar on NLP Annotation Tool for CoEs
- Mar 18th, 2024: Webinar on New Version of the Capsule
- Jun 10th, 2024: Webinar on Data Source Quality Mapping Tool
- Jul 1st, 2024: Webinar on Validating the final RAVEN mock-ups (WP6)

For each Webinar, a relevant subfolder in the shared repository of the WG on Pilot Deployment is created, containing the video recording of the Webinar as well as any slide decks or other supporting materials used during the Webinar delivery.

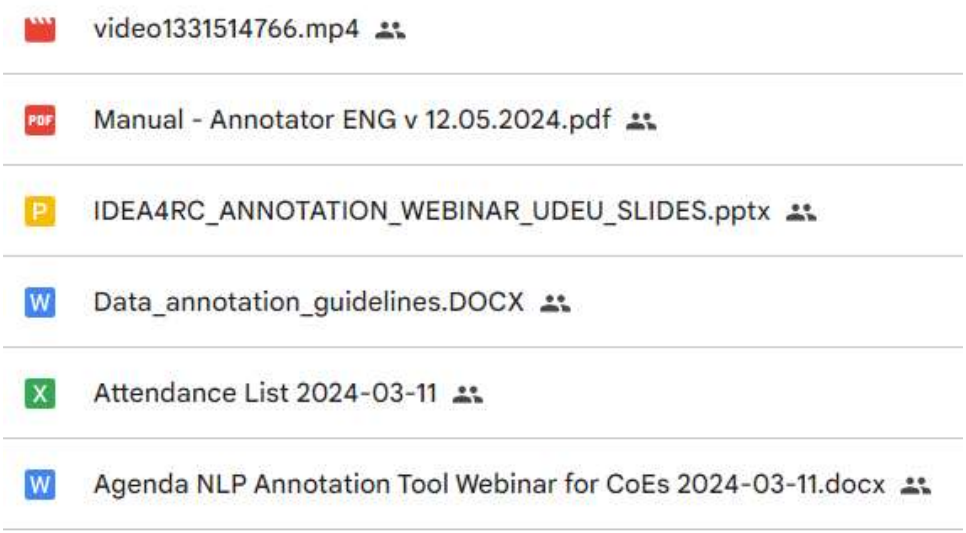

Figure 9 - Example of materials for a WG on Pilot Deployment Webinar

Webinars represent a further venue, where the project technical team and CoEs representatives can convene and discuss deployment matters, with particular emphasis on specific adjustment to the local CoE contexts, to drive subsequent activities accordingly.

5 ADAPTING TO LOCAL PILOT SPECIFICITIES

As discussed in the previous sections, adaptability is one of the mandatory requirements for the IDEA4RC Pilot Deployment planning.

With reference to activities illustrated in section 3 and at the time of this writing, CoEs specificities that require special technical adaptation fall into three categories:

- *Point of ingestion*: where the CoE datasets are to be ingested into the IDEA4RC platform (in particular, the IDEA4RC Capsule). Four possibilities are foreseen:
    - P1: through a normalized table, aligned with the IDEA4RC Data Model
    - P2: directly into the IDEA4RC Data Model
    - P3: from a FHIR representation of data
    - P4: from an OMOP representation of data
- *NLP algorithms*, to extract structured data from text data in natural language. Three possibilities are foreseen:
    - IDEA4RC: the CoE will use IDEA4RC NLP algorithms
    - Own: the CoE will use its own NLP algorithms
    - None: the CoE does not need NLP extraction
- (in case NLP extraction is used) *NLP annotation tool*, to annotate textual data to be used for training NLP algorithms. Two possibilities are foreseen:
    - CLN Tool, adapted from partner CLN software, to support IDEA4RC work
    - Own: the CoE will use its own annotation tool

The following table illustrates synoptically how CoEs have opted along the above-mentioned categories and, in addition, how IDEA4RC technical partners are expected to accordingly support each CoE, in order to adapt to each particular situation.

Efforts similar to the one illustrated in the table below might also be conduced in the future, if and as soon as additional needs for CoE customization would arise.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Table 1 – Adapting to local CoE specificities

				Technical partner support roles						
	Point of ingestion	**NLP algorithms**	**NLP annotation tool**	**UPM**	**ENG**	**UDEUSTO**	**FBK**	**CERTH**	**TNO**	**CLN**
INT	P1 (Normalized Table)	IDEA4RC	CLN	Support for capsules deployment (s)	Support for capsules deployment and raw data ETL for data ingestion (se)	Support for data quality dashboard, annotation and NLP training (dan)	NLP training and provision of data extraction algorithms (ner)	Support for governance layer integration (gi)	Support for data permit signing services (dp)	Support for Clininote annotation tool (ca)

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

				Technical partner support roles						
	Point of ingestion	**NLP algorithms**	**NLP annotation tool**	**UPM**	**ENG**	**UDEUSTO**	**FBK**	**CERTH**	**TNO**	**CLN**
CLB	P4 (OMOP)	None		Support for capsules deployment and OMOP ETL for data ingestion (so)	Support for capsules deployment (s)	Support for data quality dashboard (dq)	NLP model share (ns)	Support for governance layer integration (gi)	Support for data permit signing services (dp)	Support for local annotation tool (cl)

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

				Technical partner support roles						
	Point of ingestion	**NLP algorithms**	**NLP annotation tool**	**UPM**	**ENG**	**UDEUSTO**	**FBK**	**CERTH**	**TNO**	**CLN**
APHP	P1 (Normalized Table)	IDEA4RC	Own	Support for capsules deployment (s)	Support for capsules deployment and raw data ETL for data ingestion (se)	Support for data quality dashboard and NLP data sharing(dqs)	NLP training and provision of data extraction algorithms (ner)	Support for governance layer integration (gi)	Support for data permit signing services (dp)	Support for local annotation tool (cl)
IIS-FJD	P4 (OMOP)	IDEA4RC	Own	Support for capsules deployment and OMOP ETL for data ingestion (so)	Support for capsules deployment (s)	Support for data quality dashboard and NLP data sharing(dqs)	NLP training and provision of data extraction algorithms (ner)	Support for governance layer integration (gi)	Support for data permit signing services (dp)	Support for local annotation tool (cl)
VGR	P1 (Normalized Table)	IDEA4RC	CLN	Support for capsules deployment (s)	Support for capsules deployment and raw data ETL for data ingestion (se)	Support for data quality dashboard, annotation and NLP training (dan)	NLP training and provision of data extraction algorithms (ner)	Support for governance layer integration (gi)	Support for data permit signing services (dp)	Support for Clininote annotation tool (ca)

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

				Technical partner support roles						
	Point of ingestion	**NLP algorithms**	**NLP annotation tool**	**UPM**	**ENG**	**UDEUSTO**	**FBK**	**CERTH**	**TNO**	**CLN**
MSCI	P1 (Normalized Table)	IDEA4RC	CLN	Support for capsules deployment (s)	Support for capsules deployment and raw data ETL for data ingestion (se)	Support for data quality dashboard, annotation and NLP training (dan)	NLP training and provision of data extraction algorithms (ner)	Support for governance layer integration (gi)	Support for data permit signing services (dp)	Support for Clininote annotation tool (ca)
MUH	P1 (Normalized Table)	IDEA4RC	Own	Support for capsules deployment (s)	Support for capsules deployment and raw data ETL for data ingestion (se)	Support for data quality dashboard and NLP data sharing(dqs)	NLP training and provision of data extraction algorithms (ner)	Support for governance layer integration (gi)	Support for data permit signing services (dp)	Support for local annotation tool (cl)
OUS/NIPH	P4 (OMOP)	None		Support for capsules deployment and OMOP ETL for data ingestion (so)	Support for capsules deployment (s)	Support for data quality dashboard (dq)	NLP model share (ns)	Support for governance layer integration (gi)	Support for data permit signing services (dp)	Support for local annotation tool (cl)

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

				Technical partner support roles						
	Point of ingestion	**NLP algorithms**	**NLP annotation tool**	**UPM**	**ENG**	**UDEUSTO**	**FBK**	**CERTH**	**TNO**	**CLN**
MMCI	P4 (OMOP)	Own	Own	Support for capsules deployment and OMOP ETL for data ingestion (so)	Support for capsules deployment (s)	Support for data quality dashboard and NLP data sharing(dqs)	NLP training data and model share (nerd)	Support for governance layer integration (gi)	Support for data permit signing services (dp)	Support for local annotation tool (cl)
FPNS	P1 (Normalized Table)	IDEA4RC	CLN	Support for capsules deployment (s)	Support for capsules deployment and raw data ETL for data ingestion (se)	Support for data quality dashboard, annotation and NLP training (dan)	NLP training and provision of data extraction algorithms (ner)	Support for governance layer integration (gi)	Support for data permit signing services (dp)	Support for Clininote annotation tool (ca)
UKE	P3 (FHIR)	IDEA4RC	Own	Support for capsules deployment (s)	Support for capsules deployment and FHIR data ETL for data ingestion (sf)	Support for data quality dashboard and NLP data sharing(dqs)	NLP training data and model share (nerd)	Support for governance layer integration (gi)	Support for data permit signing services (dp)	Support for local annotation tool (cl)

Legenda:

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

s - Support for capsules deployment, for this task the support of the partner is important to speed up the process and provide customization if needed for the specific COE
so - Support for capsules deployment and OMOP ETL for data ingestion, for this task the support of the partner is important to speed up the process and provide customization of deployment if needed for the specific COE and, in addition the partner is foudamentel to debug the process of ingestion of the data into the capsule in the OMOP format.
se - Support for capsules deployment and OMOP ETL for data ingestion, for this task the support of the partner is important to speed up the process and provide customization of deployment if needed for the specific COE and, in addition the partner is foudamentel to debug the process of ingestion of the data into the capsule in the raw csv format.
dan - Support for data quality dashboard, annotation and NLP training. The partner is essential for the CoE to help and support in the activities asociated to NL and data quality, where deep debug and undestanding on related tools is mandatory.
dq - Support for data quality dashboard. The partner is essential for the CoE to help and support in the activities related to data quality, where deep debug and undestanding of the tools is mandatory.
ner - NLP training and provision of data extraction algorithms. The partner is essential for the CoE because is the only one that can understand the process for integrating and use the algorithms for structuring unstructured data.
ns - NLP model share. The partner is useful for integrating the different models used at CoE with the one developed in IDEA4RC for enhancing performance for structuring unstructered data.
nerd - NLP training data and model share. The expertise of the partner is nedded for sharing and adapt data and models for training in different context with different data models.
sf - Support for capsules deployment and FHIR ETL for data ingestion, for this task the support of the partner is important to speed up the process and provide customization of deployment if needed for the specific COE and, in addition the partner is foudamentel to debug the process of ingestion of the data into the capsule in the FHIR format.

dqs - Support for data quality dashboard and NLP data sharing. The partner is essential for the CoE to help and support in the activities related to data quality and NLP model sharing, where deep debug and undestanding of data, tools and NLP models is mandatory.
gi - Support for governance layer integration. The partner is the owner of the governance layer and its knowledge is required for customization for the specific needs of the CoE related to the governance.
ca - Support for Clininote annotation tool. The partner is the owner of the annotation tool, for this reason its knowledge is required for the specific needs of the CoE related to the annotation of unstructured data.
cl - Support for local annotation tool. The partener may help CoE in the local annotation task due to its unique knowledge in the field.

6 CONCLUSIONS

The deployment of IDEA4RC project aims to mark a significant advancement in federated research, particularly in the domain of rare cancer epidemiology. By structuring unstructured data, integrating cutting-edge NLP and federated learning technologies, and maintaining stringent privacy standards, IDEA4RC aims to set a new standard for collaborative research.

The phased approach expect that the deployment is both methodical and adaptable, allowing for iterative improvements and the accommodation of local pilot site specificities. Key milestones, such as the deployment of capsule versions with both fake and real data, underscore the project's progress toward its ultimate goal of enabling efficient, secure data sharing and research.

The success of this deployment hinges on the collaborative efforts of all stakeholders, supported by robust governance structures, continuous monitoring, and adaptable DevOps practices. As the project move forward, the focus will remain on refining these processes and ensuring that the project delivers tangible benefits to the research community and, ultimately, to patients affected by rare cancers.