



A Multimodal AI-based Toolbox and an Interoperable Health Imaging Repository for the Empowerment of Imaging Analysis related to the Diagnosis, Prediction and Follow-up of Cancer

Deliverable 5.3

INCISIVE pan-European repository of health images (final version)

WP 5 – INCISIVE pan-European repository of health images

31-07-2023

Revision 1.0

Status: Final

Grant Agreement n 952179



DOCUMENT CONTROL	
Project reference	Grant Agreement number: 952179
Document name	INCISIVE pan-European repository of health images (final version)
Work Package	WP 5
Work Package Title	INCISIVE pan-European repository of health images
Dissemination level	CO
Revision	1.0
Status	Ready for submission
Reviewers	Berta Borràs, Laia Juan, Olalla Aramburu (MDT) Sampsa Hautaniemi (UH)
Beneficiary(ies)	CERTH

Dissemination level:

PU = Public, for wide dissemination (public deliverables shall be of a professional standard in a form suitable for print or electronic publication) or CO = Confidential, limited to project participants and European Commission.

AUTHORS		
	Name	Organisation
Document leader	Kostas Votis	CERTH
Participants	Zisis Sakellariou	CERTH
	Anastasios Tzelepakis	CERTH
	Nikolaos Siopis	CERTH
	Paschalis Bizopoulos	CERTH
	Evangelos Politis	CERTH
	Giannis Aslanis	CERTH
	Dimitrios Manolakis	CERTH
	Antonios Lalas	CERTH
	Chrysostomos Symvoulidis	TIS
	Alexandra Kosvyra	AUTH
	Caroline Barelle	ED
	Pablo Mezzon	ED
	Hara Stefanou	ADAPTIT
Paris Laras	MAG	

REVISION HISTORY				
Revision	Date	Author	Organisation	Description
0.1	22/06/2023	Zisis Sakellariou	CERTH	ToC
0.2	26/06/2023	Zisis Sakellariou	CERTH	ToC Updated
0.3	03/07/2023	Alexandra Kosvyra, Chrysostomos Symvoulidis, Zisis Sakellariou, Anastasios Tzelepakis, Paschalis Bizopoulos, Nikolaos Siopis, Evangelos Politis, Giannis Aslanis, Dimitrios Manolakis Antonios Lalas, Konstantinos Votis	AUTH/TIS/CERTH	Initial content
0.4	16/07/2023	Hara Stefanou, Chrysostomos Symvoulidis, Zisis Sakellariou	ADAPTIT/TIS/CERTH	Additional content
0.5	19/07/2023	Caroline Barelle, Pablo Mezzon	ED	Additional content
0.6	21/07/2023	Chrysostomos Symvoulidis, Paris Laras, Caroline Barelle, Zisis Sakellariou	TIS/MAG/ ED/CERTH	Final input
0.7	25/07/2023	Zisis Sakellariou	CERTH	Ready for peer review
0.8	26/07/2023	Zisis Sakellariou, Caroline Barelle, Chrysostomos Symvoulidis	CERTH/ED/TIS	Addressing comments from peer review
0.9	28/07/2023	Zisis Sakellariou, Antonios Lalas, Konstantinos Votis	CERTH	Final check
1.0	28/07/2023	Konstantinos Votis	CERTH	Ready for submission

Disclaimer and statement of originality

The content of this deliverable represents the views of the authors only and is their sole responsibility; it cannot be considered to reflect the views of the European Commission and/or the Consumers, Health, Agriculture and Food Executive Agency or any other body of the European Union. The European Commission and the Agency do not accept any responsibility for use of its contents.

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made

Table of Contents

1. Introduction	8
1.1. Purpose and scope	8
1.2. Document structure	8
1.3. Relation with other deliverables	9
2. Addressing P1 review recommendations.....	10
3. Overview of the INCISIVE Repository Implementation and Updates Compared with the Second Prototype	12
3.1. Summary of the updates.....	12
4. The INCISIVE Hybrid Repository Design	14
4.1. Data preparation: quality check, de-identification and annotation	14
4.1.1. Overall description.....	14
4.1.2. Updates compared with the second prototype	15
4.1.3. Use cases considered	15
4.1.4. Functionality implementation according to the foreseen use cases	18
4.2. Hybrid data storage functionality and INCISIVE Common Data Model	25
4.2.1. Overall description.....	25
4.2.2. Updates compared to the second prototype.....	25
4.3. Data sharing schema	39
4.3.1. Overall description.....	39
4.3.2. Updates compared with the second prototype	40
4.3.3. Functionality implementation according to the foreseen use cases	40
5. INCISIVE Repository user interface.....	47
5.1. Data preparation: quality check, de-identification and annotation	47
5.1.1. DICOM De-identification and curation tool.....	47
5.1.2. Semi-automatic annotation tool	48
5.1.3. Data Integration Quality Check Tool	48
5.2. INCISIVE data search	51
5.3. INCISIVE workspaces	53
5.4. Administration.....	55
5.5. Data sharing portal	57
Registration form fields.....	60
Candidate Data Provider Form	60
Candidate Data User Form.....	60
6. Conclusions	62
References.....	63
ANNEX I – System Requirements Checklist	64
ANNEX II – Curation Script.....	67

Table of Figures

Figure 1. Final infrastructure of the INCISIVE hybrid repository	12
Figure 2. Results for the 3 rounds of user validation.....	24
Figure 3. Central Space infrastructure	31
Figure 4. Search results for a specific search.....	33
Figure 5. Auditing mechanism workflow.....	37
Figure 6. Reputation system workflow.....	38
Figure 7. Data sharing process.....	41
Figure 8. Data sharing mechanism.....	42
Figure 9. Data Provider's registration process.....	43
Figure 10. Data Use process.....	44
Figure 11. Data Use mechanism.....	45
Figure 12. Data User registration process.....	46
Figure 13. De-identification and curation tool interface.....	47
Figure 14. Semi-automatic annotation tool interface The blue area indicates the region of the lung (in this case), that has been annotated, using the brush tool.....	48
Figure 15. The login page of the DIQCT.....	50
Figure 16. The results of the newly added DICOM Validation Component.....	50
Figure 17. The results of the newly added Annotation Component.....	51
Figure 18. Performing a search in the platform.....	51
Figure 19. Search results page.....	52
Figure 20. View detailed information of a search.....	53
Figure 21. Select the creation of a Workspace from a search.....	54
Figure 22. Create a Workspace.....	54
Figure 23. List of available Workspaces.....	55
Figure 24. List of available Services.....	56
Figure 25. Create a new AI Service or a Pipeline.....	56
Figure 26. Logs inspection from the administrator.....	57
Figure 27. Home-page of the Data Sharing portal.....	58
Figure 28. Openly accessible information for DPs.....	59
Figure 29. Openly accessible information for DUs.....	59
Figure 30. Registration form for DPs.....	61
Figure 31. Registration form for DUs.....	61
Figure 32. Proposed folder structure.....	67
Figure 33. Typical example of a generated report containing the precise issues and the proposed solution.....	68
Figure 34. Report generation after a data provider uploads new data to the central node.....	69
Figure 35. Merging data corrected by the data providers with data corrected by the curation script.....	71

Table of Tables

Table 1. Addressing P1 review recommendations.....	10
Table 2. Registrations form for DP's and DU's.....	60
Table 3. Non-functional System requirements.....	64
Table 4. Functional System requirements.....	65

Abbreviations

Abbreviation	Description
AI	Artificial Intelligence
API	Application Programming Interface
CDM	Common Data Model
CS	Central Storage
D	Deliverable
DevOps	Development Operations
DICOM	Digital Imaging and Communication in Medicine
DIQCT	Data Integration Quality Check Tool
DLT	Distributed Ledger
DM	Data Management
DP	Data Provider
DU	Data User
ETL	Extract Transform Load
FAIR	Findable Accessible Interoperable Reusable
FHIR	Fast Healthcare interoperability resources
FR	Federated repository
FS	Federated Storage
GDPR	General Data Protection Regulation
HCP	Healthcare Professional
HLF	Hyperledger Fabric
M	Month
MLOps	Machine Learning Operations
No	Number
SDK	Software Development Kit
SE	Search engine
SFTP	Secure File Transfer Protocol
UC	Use Case
UI	User Interface
VM	Virtual Machine
WP	Work Package

1. Introduction

1.1. Purpose and scope

This deliverable is the third and final iteration of the WP5 outcome, referring to the design and implementation of the INCISIVE Pan-European federated repository of health images.

The tasks under this WP aim to implement functionalities that are tied to data management, such as data preparation before the data are shared, curation and storage of the data, security and accountability mechanisms.

The work presented in this deliverable complements the work that has been carried out in D5.2 [5] (second version of the NCISIVE Pan-European Federated Repository) and describes the updates performed under WP5 task in the time period since D5.2, and how they are presented in the UI. This deliverable is tied to D3.3 [3], which is the final version of the INCISIVE infrastructure, where the final architecture for each component is described.

The main updates of D5.3 in comparison with D5.2 concern the finalization of: (1) the data preparation tools, (2) the data sharing portal, (3) the functionalities for storing and searching data in the hybrid repository, and (4) updates on the UI.

1.2. Document structure

D5.3 is divided in the following sections:

Section 2 – Addressing P1 review recommendations: describes the actions taken to further address the recommendations related to the INCISIVE Pan-European Federated Repository received from the reviewers during the P1 review.

Section 3 – Overview of the INCISIVE Repository Implementation and Updates Compared with the Second Prototype: depicts briefly the INCISIVE hybrid repository architecture that serve as a basis for INCISIVE development purposes depicted in D2.5. It also highlights the updates compared with the second prototype. It focuses also on the data management functionalities of the repository: Data preparation, storage/sharing and searching.

Section 4 – The INCISIVE Hybrid Repository Design: specifies the first two layers of the INCISIVE data management according to the use-cases identified in D2.3 [1] and updated in D2.5 [2]. In particular, it addresses: (1) The data preparation functionality that consists of the data curation, de-identification and annotation processes, and data integration quality check; and (2) The data storage and sharing functionality considering the hybrid feature of prototype 2 according to interoperable and FAIR [8] principles as well as GDPR regulation [9].

Section 5 – INCISIVE Repository user interface: presents the updates in the UI of the INCISIVE platform, the set of tools provided alongside with the platform, as well as the data sharing portal.

Section 6 – Conclusion: concludes the deliverable and summarizes the main work presented in D5.3, the final iteration of the INCISIVE Pan-European repository.

1.3. Relation with other deliverables

This deliverable is the final iteration of the INCISIVE pan-European federated repository and extends the work of D5.2. The work and efforts described in this document are closely linked to work performed under other WPs, such as WP3 and deliverable D3.3, WP6 and deliverable D6.2 [6]. It is also associated with the work under WP4 and D4.3 [4], and ultimately the final architecture of the INCISIVE platform described in WP2 and D2.5, as well as, with work performed under WP8 and its deliverable D8.5 [7]. The current document is the final version that will be integrated to INCISIVE final prototype, delivered on M36.

2. Addressing P1 review recommendations

The P1 review recommendations have been addressed in deliverable D5.2. However, in this iteration, more actions have been taken by the consortium to further address them at the best way possible.

Table 1. Addressing P1 review recommendations.

ID	Review Recommendation	Work performed in D5.3	Links to D5.3
R2.2	<p>System user interfaces should be improved as soon as possible. The look and feel should be improved and branded with a clear project identity. They should be adapted to the target audience, have intuitive user navigation, feedback on user actions, familiar UI elements, attractive visuals design, easily accessible, and compliance with new design standards.</p>	<p>We updated the user interfaces in order to meet the users’ needs and make the platform overall more user-friendly using the new design standards. In more detail, the purpose of our redesigned platform is to create a simple, clear and easy-to-navigate interfaces in order to help the user.</p> <p>The colours used are in line with the logo, but they are subtle enough to aid navigation and not to create visual noise as the main goal of this site is to provide tools and informational data for the experts to use.</p> <p>The different kinds of information and functionalities are organized and then presented the same way every time, such as in cases of tables visualisations, since continuity helps the users to easily learn a system and perform actions intuitively.</p> <p>Furthermore, we enhanced certain functionalities by including tools such as sorters, filters, search fields, dropdown lists in the tables, while also providing quick actions in the homepages, which can help the user manage and filter all information and the content according to their needs and quickly perform the most frequent actions.</p> <p>The INCISIVE portal’s user interface is also further enhanced, based on comments received by actual users of the system. In addition, more functionalities are integrated in order to assist further the users by allowing to perform more actions through the platform. These include enhancements to the federated search functionality, where the users are now</p>	<p>Section 5.5: UI updates</p>

		<p>able to differentiate their queries based on the cancer type, the search results include more crucial information which should assist the users on selecting the most appropriate data, while the use of AI services and pipelines is refined to make the user better understand the scope of each service, how it is used and what are the results and the predictions it provides.</p>	
<p>R2.5</p>	<p>Smart contracts used to authenticate users and to track who has been adding or changing information in the Images or EHRs are a good start of using Blockchain technology. Smart contracts for accessing, understanding (XAI) and tracking the basis data used to form final AI models could be interesting areas if time allows. For example, see work proposed in: Nassar, M., Salah, K., Rehman, M.H. and Svetinovic, D., 2020. Blockchain for explainable and trustworthy artificial intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(1), p.e1340.</p>	<p>Using this recommendation as a starting point, to study the existing literature and propose a solution, as a consortium, we concluded that we would use the blockchain for a reputation system for the AI algorithms used in INCISIVE.</p> <p>A number of possible solutions were presented, based on existing literature, such as preventing the one-pixel attack in the XAI feature. However, such solutions would present additional requirements that the Transaction Tracker could not fulfil at this stage of implementation and at this timeframe. The next solution presented was the creation of a reputation system for the XAI algorithms, based on user satisfaction, however, such a mechanism would present an overlap with work under another WP of the project. Hence, the proposed solution is a reputation system for the AI algorithms, based on user answering a set of predefined questions, from which a score is computed and assigned to each algorithm, representing this algorithm's reputation score.</p>	<p>Section 4.2.2: Transactional auditing and accountability mechanisms</p>

3. Overview of the INCISIVE Repository Implementation and Updates Compared with the Second Prototype

The creation of a pan-European federated repository for health images and data that can be used for prompt cancer detection as well as prediction and follow-up with Machine Learning and AI algorithms, is one of the main outcomes of the INCISIVE project. The existence of such a repository can be a solution to the problem of data availability when it comes to the availability of high-quality image datasets that fulfil a set of criteria, such as following well-known principles (e.g., GDPR), while also being insightful and ready for use by the AI algorithms. The INCISIVE repository follows a hybrid approach, providing both a Central Node and Federated Nodes, as it has been defined in D5.2. The final infrastructure of the repository has already been defined in D2.5, and is shown in Figure 1.

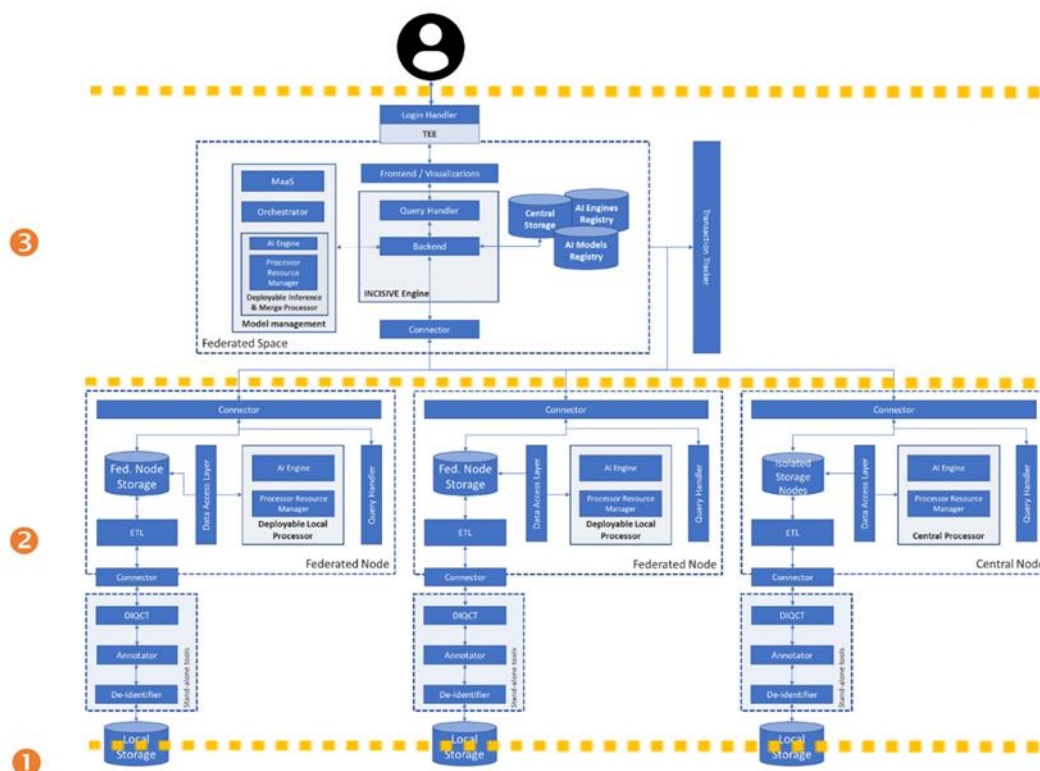


Figure 1. Final infrastructure of the INCISIVE hybrid repository

3.1. Summary of the updates

Since the infrastructure and architecture has not changed significantly since the previous iteration, it is still separated in three layers:

1-The data preparation tools: The suite of the data preparation tools, which includes the data de-identification tool, the semi-automatic annotation tool and the quality check tool, has been finalized and is ready for use, ensuring data harmonization and the proper pre-processing of data, before being shared and stored in the INCISIVE repository.

2-The Federated/Central Node layer: The Federated and Central Node layer that allow Data providers according to their needs/possibilities in terms of infrastructure, to store and share their data while keeping full control over it are at the time of writing this deliverable deployed. This layer is composed of:

- 6 federated nodes controlled by their respective data providers, deployed at disparate clinical sites
- 1 central node server currently virtualizing 9 Virtual Nodes to support the INCISIVE hybrid version available for data providers that opt for such virtual nodes

3-The Federated Space layer that is the Cloud environment that contains the centralized services required to make the INCISIVE platform work has been updated to include MLOps pipeline functionalities, as well as the capacity for local-only inference, in order to optimise the various types of workflows INCISIVE provides. In particular Apache Airflow has been replaced by the integration of Argo Workflows. This replacement is detailed and justified in D3.3 'Infrastructure & DevOps Environment Setup – Final Version'.

4. The INCISIVE Hybrid Repository Design

This chapter of the deliverable covers the main updates that concern the data preparation, data storage and data sharing framework. Since this is the last iteration of the INCISIVE pan-European federated repository of health images, the updated and the new features of each component elaborated below, are considered final and provide the complete functionality promised to be implemented for the INCISIVE project. The updates and additions refer to:

- The data preparation process, which is carried out via a set of tools
- The data sharing mechanism, that provides all the required information for potential new Data Providers and Data Users, and the mechanisms that allow them to do so
- The data storage and search functionalities, which is based on a federated approach, with both a central node and optional federated nodes, supplemented by a set of functionalities that guarantee privacy and accountability.

For consistency purposes, the same structure with the previous iteration of this deliverable is followed, which includes:

- An overall description
- A summary of the major updates performed
- The use cases taken into account
- The functionality implementation, according to the foreseen use cases

4.1. Data preparation: quality check, de-identification and annotation

4.1.1. Overall description

The data preparation process is of utmost importance, especially when this process entails sensitive data, such as medical data and images. The goal of data preparation is to create high quality datasets that can be used for the proper training of the AI algorithms, aiming to produce trustful and accurate results, that will aid healthcare professionals in their everyday workflow. Data preparation is a rather meticulous process in which data, after being processed, must retain their usability for their intended purpose of use, while ensuring user confidentiality and compliance to the GDPR.

In INCISIVE a plethora of data are utilized, such as medical data in the form of DICOM images, clinical data, in the form of templates in .xls format. While the clinical data do not contain any personal or identifying information about the user, DICOM data must undergo a number of steps and processes. These steps include data de-identification, data annotation and data quality check, for ensuring that the de-identified data have successfully been deprived of any element that could lead to patient re-identification, while following the harmonization requirements.

A supplementary step has been added to this process, concerning data curation. Data curation was already present in the data preparation steps, as it was carried out automatically

during the de-identification process and by adhering to guidelines, specifically prepared for the INCISIVE project. However, due to some deviations from the guidelines, additional effort was made to implement a service for curating data that presented inconsistencies either in the folder structures and naming conventions, or misplacements of data.

The tools developed under the umbrella of the data preparation process have been designed, implemented and updated (since their previous iterations), prioritizing the needs of the healthcare professionals, while trying to maintain the tools as easy to use as possible. The de-identification, semi-automatic annotation and quality checking tools have been implemented and are ready for use by the Data Providers, for sharing their data within the INCISIVE project and repository, while the curation service is executed internally by the technical team and notified any Data Provider whose data contain issues that require their attention and prompt actions for resolving them, via an offline process.

4.1.2. Updates compared with the second prototype

The major updates that have been implemented are related to the Data Integration Quality Check Tool (DIQCT) and the curation script. A summary of these updates follows, while a more extensive description of them can be found in the respective parts of this document.

1. Data Integration Quality Check Tool: The main updates are the addition of a new login mechanism, a DICOM image validator, that checks if the images are valid DICOM images, and a component that checks for the proper placement of annotation files.
2. Curation: This new process takes place once the data have been uploaded and checks about possible issues that stem from not strictly adhering to the provided guidelines. Issues that can be fixed automatically, such as deletion of empty folders or split of non-annotated multi-sequence series folders, are resolved immediately, whereas issues that require a specific Data Provider to take some actions, are communicated offline. This process has the form of a script written in Python and is executed periodically during the retrospective and prospective data uploading procedure.

The rest of the tools present minor updates since the Second prototype, and they are mentioned below, in the respective sections.

4.1.3. Use cases considered

Data management

UC Code	Title	Description
UC-DM-01	Data de-identification	The de-identification and curation tool accepts data in the form of medical files (DICOM 3D MRI or CT scans) and produce de-identified image files (DICOM) in which the fields that contain personal information

		such as name and date of birth, are either replaced with fake data or completely removed. This tool utilizes the de-identification protocol that was based on the NEMA de-identification suggestions.
UC-DM-02	Data annotation	The semi-automatic annotation tool accepts data in the form of medical files (DICOM/NIfTI MRI or CT scans) or slices of these files of specific regions of medical interest. It produces image annotation files (NIfTI) that consist of a mask of the specific region of medical interest on the pixel level using pretrained models or a functionality of manual annotation (brush).
UC-DM-03	Data quality check	Collected data are stored in the local premises of the user that intends to upload them into the INCISIVE platform. In order to integrate the data into the INCISIVE platform, a data quality check must be applied. The goal of the check is to identify whether the data follow the harmonization requirements reflecting, among others, the de-identification protocol applied, the inclusion of all the required information on the accompanying metadata as well as the integrity. The folder structure is also checked. A consistency check between the case metadata in the imaging (DICOM) patient data also take place. The outcome of the data quality check are reports proposing corrective actions to the user.

Storage and Search

UC Code	Title	Description
UC-FS-01	Data Sharing	The user sharing data, is granted access for this scope. The data can be then indexed in the INCISIVE hybrid repository, through the INCISIVE data sharing mechanism.
UC-FS-02	Collection of Data upon a specific query /request	The user (with the assumption of being granted the corresponding access rights) has the ability to perform

		a search for specific datasets on the INCISIVE repository, based on specific criteria.
UC-FS-03	Transactional auditing and accountability	The transactional auditing and accountability mechanism allows at least and among others, the transactions tracking for the data sharing and re-use functions, ensuring accountability.

Security

UC Code	Title	Description
UC-SU-01	Logging and auditing blockchain-based mechanism	Authorized users perform any transactional data activities in the federated repository. These activities include activities such as data submission, access. Critical actions are recorded in a trusted Distributed Ledger (DLT), using blockchain technology. This involves creating a transaction log after each critical action (e.g., data access), containing action data and/or metadata, which is then broadcasted among all involved peers. This log is written in the blockchain in an immutable way and can be used as digital evidence, to resolve future conflicts.
UC-SU-02	Security against attacks within the operating system	An operation containing sensitive data needs to be done by an authorized user. In order to avoid any unauthorized access to sensitive data/images, hardware security extensions will be utilized. These extensions aim to define protected areas of the memory, called enclaves. In those enclaves the data is encrypted and can only be accessed by certain processes.
UC-SU-03	Data access control mechanism	A user requests to obtain access to specific data/images. In order to avoid any unauthorized access to sensitive data/images, an appropriate access control mechanism is utilized.

Administration

UC Code	Title	Description
UC-ADM-01	User Registration	In order for a user to consume the services of data sharing and re-use, storage and search s/he should be registered in the INCISIVE platform.
UC-ADM-04	Exit rights from the platform (right to be forgotten)	The registered user does not want any more to share data or make queries to the INCISIVE platform. He can exit the platform thanks to an exit mechanism. The exit is recorded.

4.1.4. Functionality implementation according to the foreseen use cases

UC Code	Ref. System	Title
UC-DM-01	Data de-identification	DICOM de-identification and curation tool

Prototype functionality related to the UC

The de-identification and curation tool developed under T5.2 offers a variety of features, and useful functionalities that allows users to de-identify DICOM images and curate the results, by creating a folder structure and appropriate folder names.

Although, the aforementioned tool guarantees curation of the de-identified files, the technical team developed a curation script, that takes care of curation issues that came up during data preparation and were crucial for the uniformity of data and interoperability of the different tools and algorithms. The curation script takes care of most problems such as folder structure and naming conventions. The curation script cannot handle cases where multi-sequence Series folders are present and there is no clear correspondence between NIFTI annotation files and the DICOM files within these Series.

Technical description

DICOM de-identification tool is a web-based client-side application and is used to de-identify any kind of DICOM image, by obscuring or removing the DICOM image metadata, based on the NEMA PS 3.15 protocol [10]. Each metadata field is handled according to what is foreseen, by the protocol by applying the appropriate function, however, the user is given the ability to implement specific options, once again described in the protocol that is followed. Moreover, users can manually change some fields and give them the desired value in a single image, and then apply this change automatically to all other related images. Useful new features, such as privacy, usability, duplication removal, were introduced, in

comparison with the 3rd party tool used before (CTP Anonymizer). These features have been thoroughly described in the previous version of this deliverable (D5.2).

The programming language used is JavaScript, a language suitable for implementing web-based applications, as well as HTML and CSS for the implementation of the UI.

The curation script, a Python-based tool developed for the INCISIVE project, is used in the data curation process. This script encompasses a series of automated procedures designed to address various issues associated with the retrospective and prospective data collection and uploading. Its primary objective is to ensure the accuracy, reliability, and integrity of the dataset. The script incorporates functionalities such as detecting and resolving discrepancies in the folder structure, identifying and rectifying instances of annotation files located outside of Series folders, removing empty Series folders, and reporting discrepancies between the total number of DICOM files and NIFTI slices. Additionally, the script documents any unsuccessful attempts to read DICOM and NIFTI files, facilitating future inspection and potential re-upload. In cases where automatic correction is not feasible, the script generates detailed reports that document all issues pertaining to the examined data and propose appropriate solutions. These reports are then communicated to the responsible data providers, who undertake manual inspection, correction, and re-upload of the affected data. Furthermore, the curation script is employed to merge the automatically corrected data with the manually rectified and re-uploaded data, creating a unified and refined dataset. Through its systematic execution, the curation script serves as a critical tool in ensuring the quality, consistency, and usability of the curated data for subsequent analysis, AI development, and medical research endeavours. A detailed description for the curation script and its functionalities can be found in “ANNEX II – Curation Script”.

Dependencies / interfaces

The DICOM de-identification and curation tool has its own interface, as a web application, while the sole requirement is for the user to have a modern browser and internet connection.

Since the curation script is executed periodically, and only until the data sharing is completed, it has no interface and dependencies from the final users.

Implementation description and status

Since, the tool was released a number of actions were taken and implemented behind the scenes, such as dependency management, software maintainability and continuous integration and development (CI/CD). Any bugs reported or discovered by the development team have been fixed. The DICOM de-identification and curation tool has been implemented, is fully functional and can be accessed by anyone that knows the link, or through the Data Sharing Portal.

The curation script has been implemented as an independent Python-based script and can be accessed in the central repository of the INCISIVE. Any bugs reported or discovered by the development team have been fixed.

User needs considered

The DICOM de-identification and curation tool allows for de-identifying and curating DICOM images, while providing different levels of privacy and ease of use. The curation script satisfies the needs for creating uniform datasets regarding the folder structure, naming conventions and either correcting them or produce reports, that are forwarded to DP's, with issues that have to be corrected manually.

Evaluation procedure before integration in the whole platform

The tool has been tested by the DP's and is ready for use. The curation script was tested by executing it on duplicated medical data that presented different issues and inspected the results by checking if the issues had been resolved. This process was performed internally by the development team, with assistance of other members of the consortium.

Intended procedure for users' validation

Questionnaires were circulated for user validation, where users rated the functionalities and ease of use of the new tool and compared it with the CTP Anonymizer. There was no need of user validation for the curation script.

UC Code	Ref. System	Title
UC-DM-02	Data annotation	Semi-automatic annotation tool

Prototype functionality related to the UC

The semi-automatic annotation tool that was developed under T5.2, is meant to help healthcare professionals to annotate medical images whether they are in the DICOM, NIFTI or CT scans format, in order to mark the region of interest, either performing manual annotation or in a semi-automatic way.

Technical description

This tool is a web-based client-side application, that has been developed using well established technologies and programming languages and frameworks, such as JavaScript, for the back-end implementation, and HTML and CSS for the UI. All of these technologies are appropriate for developing such an application, while ensuring privacy and a user-friendly environment.

Dependencies / interfaces

The semi-automatic annotation tool requires only a modern browser and an internet connection and has its own interface.

Implementation description and status

This tool was available since the previous version of this deliverable (D5.2) and since then minor improvements have taken place. Development and integration were continuous throughout this period and any issues encountered while using it have been resolved. The current version of the tool is fully implemented and is accessible to anyone with the link or through the Data Sharing Portal.

User needs considered

The semi-automatic tool has taken into account the needs both for manual annotations, useful for correcting existing annotations, and annotation in a semi-automatic, using pretrained models.

Evaluation procedure before integration in the whole platform

The semi-automatic annotation tool has been tested and evaluated and is ready for use.

Intended procedure for users' validation

Questionnaires were circulated through which users answered questions and provided ratings about the functionalities and ease of use of the tool.

UC Code	Ref. System	Title
UC-DM-03	Data Management (DM)	Data Integration Quality Check

Prototype functionality related to the UC

This tool is part of the data preparation process which takes place prior to data upload and is used in the local premises of the user.

Technical description

The tool is developed as a rule-based quality check. Its main purpose is to check whether the data collection requirements are followed and inform the user of potential actions that must be taken to ensure the quality of the data prior to the uploading. Moreover, the tool is extensible, the logic on checking the requirements is not hard-coded, but it is introduced from a knowledge base (specific templates, structures, anonymization protocol). The check is performed in 4 levels, clinical metadata check, images-template consistency check, de-identification protocol check, and analysis requirements check. The tool also performs a case completeness report. Its core functions include:

- Load Imaging data
- Load clinical data
- Clinical Metadata Integrity Check
- Template Structure
- Patient codification
- Timing Integrity
- Clinical data validity check
- Template-Image Consistency and Renaming
- Anonymization Protocol Check
- Analysis Requirements check
- DICOM validity check
- Annotation position and matching check
- Case Completeness measure

The tool does not intervene in the dataset in any other way except for renaming the images folders' names. Instead, it produces 4 different reports, one of each component described below. The four first reports are informative and include error messages. These messages inform the user about the issues that the data collector needs to take action to revise the data structure. The fifth is a report summarizing the dataset and accompanies the data in the data repository.

Dependencies / interfaces

The DIQCT was developed in two programming languages, R & Python and were integrated in one pipeline. This tool is related to work done prior to uploading, so it runs on the client side. The tool is implemented as Docker Image: the pipeline along with all the dependencies was built in a docker container available to all members of the consortium. This way, the pipeline can be executed in all project sites, at the local or central level.

The web interface, that presents the exposed services in a graphical way, in the client-side version, was built in R-Shiny server. The server-side version has already been implemented for the temporary infrastructure and will also run dockerized in the INCISIVE central infrastructure to produce reports on data quality.

Implementation description and status

All components of the DIQCT are implemented and incorporated in the web application.

User needs considered

This tool assists the user in correcting the dataset and bring it to a homogenized form prior to upload, by providing reports suggesting corrective actions.

Evaluation procedure before integration in the whole platform

The third and final version of the tool has been evaluated internally by the development group and externally through a testing/evaluation period with specific data providers. The round of internal validation was performed with mock-up data provided by the project partners as example cases. The feedback of the external validation was used to improve the functionality and visualization of the tool. During the internal and external validation process some issues were also identified and corrected by the team by updating the tool.

Intended procedure for users' validation

A third and final version of the user experience evaluation questionnaire has been developed and circulated through the user referring to the third version of the tool. The questionnaire explores the features of Attractiveness, Perspicuity, Efficiency and Dependability of the tool. The results of the user experience validation, for the 3 rounds of validation (each one for the 3 major versions of the tool) are depicted in Figure 2.

For the first round, a total of 9 partners participated in the survey. 6 of the participants belong to technical teams while 3 of them are medical experts. 4 of them are using the dockerized version while the other 5 are using the executable version. Finally, 7 of them are using Windows operating systems, 1 is using MAC and 1 of them uses Linux. The first part of the figure below depicts the answers to the user experience part of the questionnaire for the 4 distinct categories. The values represented for each feature are the mean value of the votes in the range 1 to 5. In the attractiveness category, it is obvious that the tool may not be so user-friendly, but the general opinion is positive. From the perspicuity category, it is concluded that the tool is easy to use and learn, understandable but not so clear. In terms of efficiency, we can say that the tool is efficient and practical but not so fast. Finally, the tool meets the expectations of the user, and it is considered supportive, although the users do not recognize the value of this tool.

For the second round, a total of 6 partners participated in the survey. 2 of the participants belong to technical teams while 4 of them are medical experts. All of them are using a dockerized version with a user interface. Finally, 5 of them are using Windows operating systems and 1 is using MAC. The second part of the figure below depicts the answers to the user experience part of the questionnaire for the 4 distinct categories. The values represented for each feature are the mean value of the votes in the range 1 to 5. In the attractiveness category, it is obvious that the tool may not be so user-friendly, and this attribute increased slightly with the user interface, but the general opinion is positive. From the perspicuity category, it is concluded that the tool is easy to use and learn, understandable but not so clear. In terms of efficiency, we can say that the tool is efficient

and practical but not so fast. Finally, the tool meets the expectations of the user, and it is considered supportive, although the users do not recognize the value of this tool. For the third round, a total of 7 partners participated in the survey. 3 of the participants belong to technical teams while 4 of them are medical experts. All of them are using a dockerized version with a user interface. Finally, 5 of them are using Windows operating systems, 1 is using MAC and 1 of them uses Linux. The third part of the figure below depicts the answers to the user experience part of the questionnaire for the 4 distinct categories. The values represented for each feature are the mean value of the votes in the range 1 to 5. In the attractiveness category, the 'enjoyable' and 'friendly' aspects have increased, as was expected with the improvement of error reporting and visualizations. In the final version the tool seems to be more understandable and practical, while we cannot notice a small decrease in efficiency, due to the addition of more components. Finally, the tool is supportive due to the opinion of the users, however the users do not recognize the value of the tool, since the 'valuable' aspect has decreased in this third round.

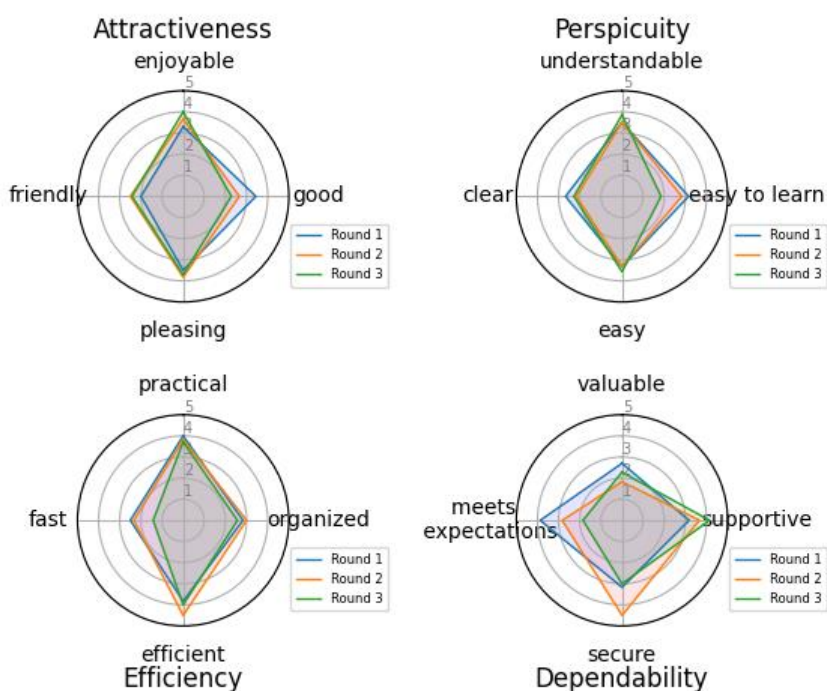


Figure 2. Results for the 3 rounds of user validation.

4.2. Hybrid data storage functionality and INCISIVE Common Data Model

4.2.1. Overall description

The functionality of hybrid data storage facilitates the archiving and distribution of both retrospective and prospective data in a hybrid style. This can be either in a federated node, which is specifically at the point of their creation such as each healthcare site, or in a central node that could essentially be a Virtual Machine (VM) provided in the Federated environment after processing through the data preparation tool mentioned in section 4.1.

The data sharing framework currently in development takes into account GDPR principles to ensure secure data management that aligns with ethical, legal, and privacy standards. We also pay close attention to data interoperability issues and the FAIR principles during the implementation of the INCISIVE Common Data Model.

4.2.2. Updates compared to the second prototype

The major updates in comparison with the second prototype are briefly presented below, while an exhaustive description for each one can be found in their respective sections in this deliverable. The main updates are:

- Hybrid repository: Central Storage has become available and data from the Temporary repository has been transferred. Minor updates have been performed in the Federated storage as well.
- Search engine: The filters for the Search Engine have been expanded, however it is still a work in progress.
- Transaction Tracker: The accountability and auditing mechanism have been completed and integrated with the necessary components (INCISIVE platform and ETL tool instances) along the required UI. Every critical action is recorded in the blockchain, including data sharing, while different users with a level of administrative rights can inspect the logs. A reputation system for the AI algorithm has been implemented, based on user input to a predefined set of questions, related to their level of satisfaction of the AI produced results. This reputation system can function as an indicator of the performance of an AI model, as well as its usability in the healthcare professionals' everyday workflow.

INCISIVE Common Data Model

UC Code	Ref. System	Title
UC-FS-01	Federated storage (FS)	Sharing of Data

UC-DM-02	Data Management (DM): INCISIVE Platform	Data foundational, structural and semantic interoperability: Common data model
----------	--	--

Prototype functionality related to the UC

The capability still aligns with the one outlined in Deliverables 5.1 and 5.2, which refers to the previous versions of the prototype. The tool or process involved facilitates data sharing following the preparation of data with the data preparation tool, which includes de-identification, quality control, annotation, and curation. Once a data provider gains access to the INCISIVE platform, data can be brought into the federated node and indexed.

In the data provider's environment, an Extract, Transform, Load (ETL) procedure harmonizes the data according to a Common Data Model, promoting interoperability and enabling federated AI. The data undergoes processing locally and doesn't need to leave the Data Provider's premises for either processing or ingestion into local storage.

This step forms the foundation of incorporating the FAIR principles into the INCISIVE platform. The INCISIVE Common Data Model leverages open standards based on HL7 FHIR and DICOM for clinical data and image data exchange, respectively, and LOINC and SNOMED CT for preserving the semantic meaning of all standardized data. The rationale for employing these interoperability standards is found in Deliverable D5.1. The FHIR Server and PACS facilitate data storage and enable its partial (metadata) or full query, ensuring:

- Data findability
- Data accessibility
- Data interoperability
- Data reusability

When the ETL procedure is finalised, the data are indexed in the INCISIVE federated data storage repository.

Technical description

A compilation of variables to be acquired for each kind of cancer (colorectal, lung, prostate, breast) has been created, grounded on the inclusion and exclusion parameters established in consultation with clinical professionals. These variables are correspondingly mapped to SNOMED CT and LOINC for clinical data, and a unique HL7 FHIR message for each type of cancer has been conceived to transfer all these variables, encoded with SNOMED CT and LOINC.

Data pertaining to cancer is gathered using a Microsoft Excel template, which is subsequently accessed by an ETL tool. This tool interprets the information and utilizes the XML formatted HL7 FHIR messages to POST the clinical data to the FHIR Server. Each HL7 FHIR message is dispatched to the corresponding FHIR server, gradually loading all data present in the template to the FHIR server. The FHIR server then ingests and orchestrates the uniform datasets within the INCISIVE infrastructure.

When it comes to images, INCISIVE follows a comparable procedure using a different ETL tool that interprets DICOM files located in a directory and subsequently transmits them to the corresponding PACS. The PACS system ingests and facilitates the management of the DICOM images within the INCISIVE infrastructure.

The components of both the ETL-FHIR Server and the ETL-PACS are provided in the form of Docker containers.

Dependencies / interfaces

Data preparation functionality

Implementation description and status

- PACS and FHIR server are functional.
- The ETL tool is equipped to process Excel files corresponding to the four types of cancer (colorectal, breast, lung, and prostate), as well as their associated DICOM images and NifTI files.

User needs considered

As in the previous version of the prototype, the following users' needs are considered:

FR1- Availability of well structured, defined and categorised data within the repository. (harmonised and standardised to make data interoperable).

FR2- Availability of large number of images with different modalities (multimodality).

FR3- Availability of raw data for each dataset.

FR4- Availability and indexation of high-quality annotated images.

FR6- Availability of demographic data, imaging data, laboratory results and biopsy results.

FR7- Linking demographic, clinical information, histopathological information and any additional health data with images.

FR8- Availability of genomic data.

FR9- Availability of time points of all medical and imaging data.

FR10- Availability of training datasets that are representative.

FR16- Ability for users to upload, search and download easily the available data.

FR12- Linking the datasets with metadata or explainable tags.

Evaluation procedure before integration in the whole platform

An end-to-end test is carried out using retrospective data from the temporary infrastructure to evaluate the relevance of the INCISIVE CDM.

Intended procedure for users' validation

The intended procedure for users’ validation of the quality of the foundational, structural and semantic interoperability of INCISIVE will be part of the usability survey intended to be implemented in September 2023 before the pilot phase. Its outcomes will enable slight refinements before piloting.

The procedure will be a mix of cognitive walkthrough method, thinking aloud, standardised usability questionnaires and a list of assessment criteria to fill.

A sample of data providers should implement a succession of tasks necessary to upload data in the federated infrastructure in order to give their opinion and thoughts about the completion of the tasks and the system. The sequence of tasks should be as follow:

- **FR Task 1** - The data provider requests access to the INCISIVE platform
- **FR Task 2** - After the data provider is granted access to the INCISIVE platform, the data provider registers themselves on the data provider’s administration page (Name, Country, URL, Contact name, logo, datasets for cancer type, information, esthesis id)
- **FR Task 3** - After the data provider is registered in the INCISIVE platform, she/he upload a dataset on the federated repository
- **FR Task 4** – After the data provider has uploaded a data a dataset on the federated repository she/he uses the functionality of the federated repository to update the dataset uploaded (cancel, delete, save modalities of the data provider administration page)
- **FR Task 5** - After the dataset has been updated, the data provider uses the functionality of the federated repository to delete the dataset updated (cancel, delete, save modalities of the data provider administration page)

INCISIVE Federated Storage

UC Code	Ref. System	Title
UC-FS-01	Federated Storage (FS)	Data Sharing
Prototype functionality related to the UC		
<p>The present version allows Data Provider (that opt for) from within their organization’s premises to host the INCISIVE Federated Storage component, which connects with the INCISIVE infrastructure. In this way, the Federated Storage in coordination with other components makes available the data present on the local machine and enables the functionalities requiring data storage in the Data Sharing use case. Currently, 6 federated nodes can be controlled by their respective data providers.</p>		
Technical description		

Every Federated Node, managed by its Data Provider, is equipped with the Federated Storage system, which is implemented within a DP's infrastructure by Maggioli. This Federated Node accommodates the data intended for sharing within its storage volumes, and the necessary components to execute the Data Sharing use case are housed within its operating system. The data, which resides on the local machine, is accessed via the ETL mechanisms inherent in the INCISIVE Common Data Model and is stored within its FHIR and PACS components. The system in place unifies these storage components with additional infrastructure components located either within the local or external infrastructures.

Dependencies / interfaces

INCISIVE Common Data Model toolbox.

Kubernetes

KubeEdge (EdgeCore, CloudCore, EdgeMesh)

Implementation description and status

Integration of INCISIVE Federated Nodes using local KubeEdge deployments has been completed. Six Data Provider Federated Nodes have been connected to the INCISIVE platform: AUTH, GOC, UOA, DISBA, UNS & HCS. Each Data Provider's Federated Node Storage data has been made available for Data Sharing.

User needs considered

FR1- Availability of well structured, defined and categorised data within the repository (harmonised and standardised to make data interoperable).

FR2- Availability of large number of images with different modalities (multimodality).

FR3 -Availability of raw data for each dataset.

FR4 -Availability of high-quality annotated images.

FR6 -Availability of demographic data, imaging data, laboratory results and biopsy results.

FR9 -Availability of time points of all medical and imaging data

FR10 -Availability of training datasets that are representative

FR16 -Ability for users to easily upload, search and download the available data.

Evaluation procedure before integration in the whole platform

The component configurations of the Federated Node Storage are manually tested by Maggioli DevOps engineers to validate their functionality, this proceeds in coordination with the developers of each software component.

Once the desired, functional configuration is established, it is used to enable automatic management of deployment and integration of each component of the Federated Node Storage and its connection with the platform

Intended procedure for users' validation

The intended procedure for users' validation of the quality of the foundational, structural and semantic interoperability of INCISIVE will be part of the usability survey intended to be

implemented targeting the second prototype planned to be finalised by M30. Its outcomes will enable further refinement for the last iteration prototype (M36).

A series of steps to allow deployment and integration of the Federated Node Storage has been provided to Data Providers through the “INCISIVE - Node Procurement & Installation Guidelines” document, which was developed under the efforts of T6.3 to produce training material for the INCISIVE repository. This document serves also as a basis to provide the Data providers of the pilot phase with a complete and user friendly “INCISIVE - Training Manual”.

INCISIVE Central Storage

UC Code	Ref. System	Title
UC-FS-02	Central Storage (CS)	Data Sharing

Prototype functionality related to the UC

This final version allows each Data Provider to provide their data securely to a central point of storage to enable sharing of their data through a centralized repository. This central repository is made available for Data Sharing and allows management access to all data providers.

Additionally, the Data Repository is made available through a Control Virtual Machine and its deployed software to INCISIVE AI Engineers and other authorized project persons to facilitate Data Sharing as well as additional functionality for Validation purposes.

The central node server currently virtualizes 9 Virtual Nodes to support the INCISIVE hybrid version.

Technical description

The system is designed to hold a large amount of data within a single repository, which is created from the individual data contributions from each Data Provider. For effective Data Sharing and management, each Data Provider's data is stored separately within the Central Storage, which is specifically designed to house the data of a given Data Provider. The Central Storage is housed within a Central Node that mimics the functionalities and components of a Federated Node, but it's virtualized within the same infrastructure. This way, the INCISIVE Central Storage uses existing components to support the Central Storage function. To ensure each Data Provider can manage their data, SFTP access is granted, allowing them to maintain direct control over their data.

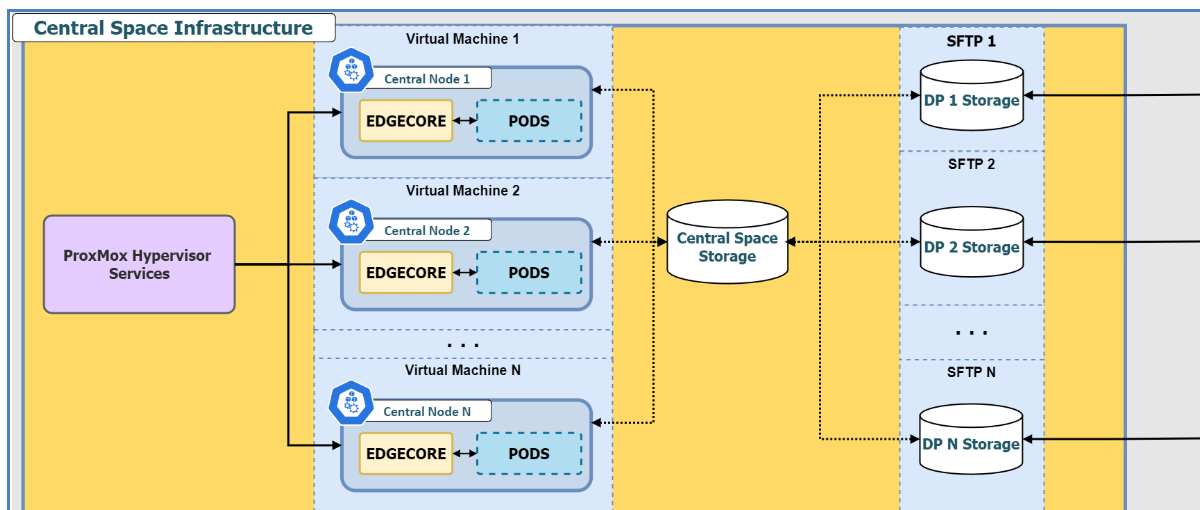


Figure 3. Central Space infrastructure.

Moreover, a Control Virtual Machine (VM) is given root access to the Data Repository. This Control VM provides direct access for Data Sharing in project activities and hosts software components for validation tasks in data and AI-related project efforts.

The infrastructure that hosts each Virtual Central Node and its related Central Storage includes a dedicated server and a backup server. These are hosted by IPHost in Haidari, Attiki, in Greece, and managed by Maggioli. They use strong hardware and fast connections to allow efficient data transfer and backup, increasing the system's overall reliability and resilience.

Dependencies / interfaces

- INCISIVE Common Data Model toolbox
- Kubernetes
- KubeEdge (EdgeCore, CloudCore, EdgeMesh)
- SFTP
- Proxmox
- ControlVM – SSH, SFTP, Python, JupyterHub Server, Anaconda

Implementation description and status

- Implementation of this system is completed.
- The Data Repository is available.
- The Control VM as well as the SFTP Data Provider data administration is enabled.
- The JupyterHub notebook server is deployed and available, credentials have been provided to authorized INCISIVE partners

User needs considered

FR1- Availability of well structured, defined and categorised data within the repository (harmonised and standardised to make data interoperable).

FR2- Availability of large number of images with different modalities (multimodality).
FR3 -Availability of raw data for each dataset.
FR4 -Availability of high-quality annotated images.
FR6 -Availability of demographic data, imaging data, laboratory results and biopsy results.
FR9 -Availability of time points of all medical and imaging data
FR10 -Availability of training datasets that are representative
FR14- Availability of a toolbox with pre-processing algorithms (tools) within the repository, for example segmentation tool for images, data harmonisation tool etc.
FR16 -Ability for users to upload, search and download easily the available data.

Evaluation procedure before integration in the whole platform

The data itself is stored in physical disk, while data access systems are virtualized within the Central Storage systems.
Each Central Storage instance is automatically validated using the deployment management tools of Kubernetes and ArgoCD.

Intended procedure for users' validation

The intended procedure for users' validation of the quality of the foundational, structural and semantic interoperability of INCISIVE will be part of the usability survey intended to be implemented beginning of September 2023 before the pilot phase.

INCISIVE Search Engine

UC Code	Ref. System	Title
UC-FS-02	Federated storage (FS)	Collection of Data upon a specific query /request
UC-DM-04	Data Management (DM)	Appropriate Dataset Selection

Prototype functionality related to the UC

The INCISIVE Search Engine enables retrospective and prospective data stored in the INCISIVE Hybrid Repository to be searched upon distributed queries in order to proceed to federated AI model development and training. Users can submit a federated search on all the INCISIVE Nodes whether federated or central, based on filters – parameters in order to find cases that match the criteria they have selected.

The filters implemented in this final version of the prototype taking into consideration users' needs extracted in different workshops are the following:

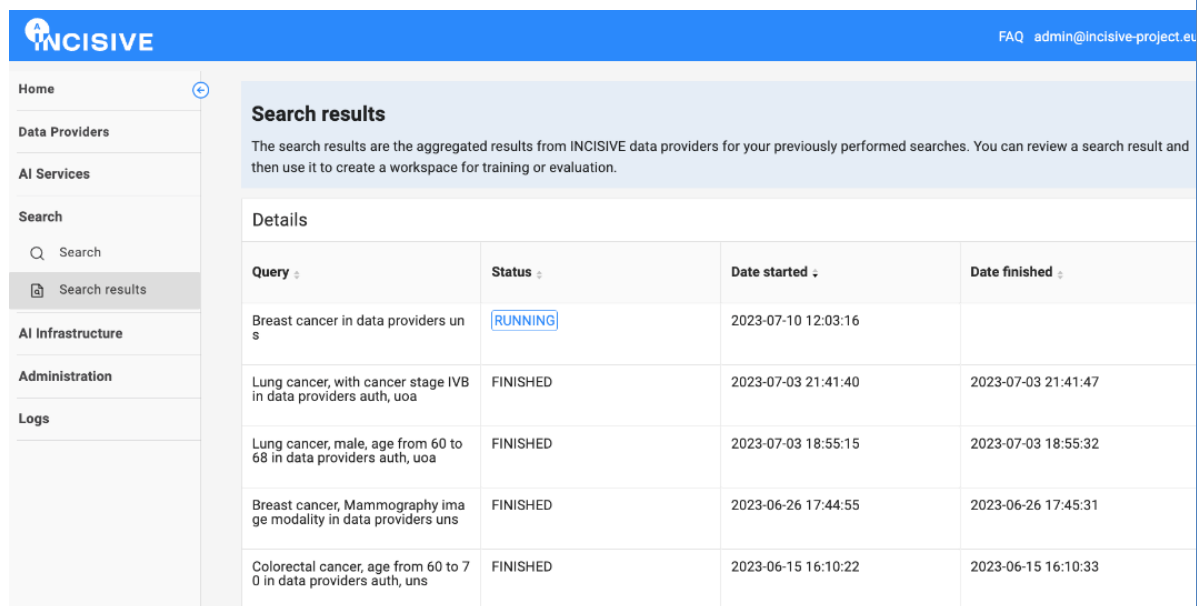
- Gender
- Age range
- Cancer type
- Cancer stage

- Months of observation
- Image modality
- Treatment/therapy
- Data Providers/Research group
- Dataset country of origin
- Datasets with full cases
- Genomic data available

The implementation of the filter related to “Timepoints Label” is still ongoing considering metadata availability (FR9- Availability of time points of all medical and imaging data.). Should a need for other filters appear before the usability study in September 2023, it will be considered upon demand and feasibility.

Technical description

A search operation is structured as an asynchronous job due to the necessity to execute across each chosen Data Provider. During this process, the web page will display the "RUNNING" status to indicate that the search is in progress. Upon completion of all Data Provider processing, the job transitions to a "FINISHED" status, as depicted in the following image. At this point, the user is able to access and review the results of their search (Figure 4).



The screenshot shows the INCISIVE web interface. The top navigation bar includes the INCISIVE logo and a contact link (FAQ admin@incisive-project.eu). A left sidebar contains navigation options: Home, Data Providers, AI Services, Search, AI Infrastructure, Administration, and Logs. The main content area is titled 'Search results' and contains a table of search details.

Query	Status	Date started	Date finished
Breast cancer in data providers un	RUNNING	2023-07-10 12:03:16	
Lung cancer, with cancer stage IVB in data providers auth, uoa	FINISHED	2023-07-03 21:41:40	2023-07-03 21:41:47
Lung cancer, male, age from 60 to 68 in data providers auth, uoa	FINISHED	2023-07-03 18:55:15	2023-07-03 18:55:32
Breast cancer, Mammography image modality in data providers uns	FINISHED	2023-06-26 17:44:55	2023-06-26 17:45:31
Colorectal cancer, age from 60 to 70 in data providers auth, uns	FINISHED	2023-06-15 16:10:22	2023-06-15 16:10:33

Figure 4. Search results for a specific search.

Other figures of the UI interface including the search and search result page are presented in section 5.

Dependencies / interfaces

The INCISIVE search engine is offered to the users via the INCISIVE UI interface. A specific page is available to the users to make their query of interest if they have successfully login to the INCISIVE platform. Once the query is launched, the results are returned to the users on the search result page of the INCISIVE UI interface. Figures of the UI interface including the search and search result page are presented in section 5.

Implementation description and status

The functionality has been updated from the previous version to address emerging needs, specifically the upgrades related to AI Services integrations. It's an automated command-line process that runs prior to the initiation of the AI process.

User needs considered

FR1 - Availability of well structured, defined and categorized data within the repository (harmonized and standardized to make data interoperable).
FR2 - Availability of large number of images with different modalities (multimodality).
FR4 - Availability of high-quality annotated images.
FR6 - Availability of demographic data, imaging data, laboratory results and biopsy results.
FR7 - Linking demographic, clinical information, histopathological information and any additional health data with images.
FR8 - Availability of genomic data.
FR9 - Availability of time points of all medical and imaging data.
FR12 - Linking the datasets with metadata or explainable tags.
FR15- Flexibility of having simple search functionality and advanced search functionality (construct complex Boolean queries).

Evaluation procedure before integration in the whole platform

The search engine including the user interface are frequently evaluated internally by the development group and externally through different workshops (e.g., Belgrade, November 2022).

Intended procedure for users' validation

The search engine enables retrospective and prospective data stored in the federated repository to be searched upon distributed queries in order to proceed to federated AI model development and training. Its validation, will be part of the usability survey (mix of cognitive walkthrough method, thinking aloud, standardized usability questionnaires and a list of assessment criteria to fill) that is intended to be implemented in September 2023 before the pilot phase and will enable slight refinements if needed. For this usability study, data users (HCP as AI services user) will only have to implement a succession of tasks in particular those pertaining to data searching as follow:

- **SE Task 1:** The data user logs into the INCISIVE platform
- **SE Task 2:** The data user accesses the search engine

- **SE Task 3:** The data user makes a search in the search page of the search engine interface for each of the 4 cancer types according to the proposed modalities: gender, age range, cancer type, cancer stage, months of observation, image modality (CT, Histopathology, MRI, PET-CT), treatment therapy (Chemo-immunotherapy, Chemo-radiotherapy, chemotherapy, Post-treatment surgery, Radiation therapy, surgery), data provider, dataset country of origin, data set with full cases, Genomic data available
- **SE Task 4:** The data user visualises the result of their search provided as a comprehensive list of the ongoing searches as well as the finished ones in the search result page of the search engine interface and where appropriate select any of the available searches in order to view detailed information related to the results and proceed to their next activities.

Transactional auditing and accountability mechanisms

UC Code	Ref. System	Title
UC-FS-03	Federated storage (FS)	Transactional auditing and accountability
UC-SU-01	Security	Logging and auditing blockchain-based mechanism
UC-SU-03	Security	Data access control mechanism

Prototype functionality related to the UC

It consists of the tools/processes that allow authorized users to perform any transactional data activities in the federated repository and to keep track of the transaction they have made, with the data provenance, the purpose of sharing, accountability. It also provides the ability for authorized users to monitor the activities performed in the platform.

Technical description

INCISIVE is a project where security, transparency and trust are significant, especially when it comes to use of medical data. The use of a distributed ledger technology (DLT), implemented via a blockchain network, and in our case a permissioned blockchain, satisfies the above requirements. The Transaction Tracker is a component is based on the Hyperledger Fabric (HLF), and is used to immutably record the fingerprints (hashes) of the most critical actions on the ledger, while the actual information of the transaction (log) along with the hash are stored in a MongoDB database, that allows for more complex queries,

when it comes to auditing purposes. The tracking functionality of the Transaction Tracker has been extensively described in previous deliverables (D5.2, D3.2).

The auditing mechanism utilizes both the blockchain, and the Smart Contracts that reside there, and the external MongoDB database. The Smart Contracts are used for verifying the identity of the user while checking if the user has the necessary rights to perform specific queries on the MongoDB, with the goal of retrieving the logs based on specific criteria and use cases, enforcing an access control mechanism. The use cases have been defined and they include monitoring the activity of all users, or for a specific user, both from the platform administrator, and each organization’s administrator. However, the latter can only perform this audit control for users under their organization. Moreover, the organization administrators and the medical professionals, under a specific organization, can check the logs based on data that were shared by this organization.

However, the use of the blockchain network has been expanded with the addition of a reputation system about the AI algorithms that implemented the services offered by the INCISIVE platform. A reputation score is attached to each of the AI algorithms, which is expressed as a mean score, coming from a voting procedure by the medical professionals. The reputation score is stored on-chain, thus providing trust and traceability on how these scores change, as only authorised users are able to invoke the underlying Smart Contract, responsible for computing and storing the reputation scores.

Dependencies / interfaces

The logging mechanism of the Transaction Tracker has no interface, since the tracking of the different actions takes place behind the scenes, via an API with multiple services exposed, each one representing a different action. On the other hand, for the auditing mechanism, an interface is provided through the INCISIVE platform only to the authorized users.

For the reputation system, although already implemented, the integration with the platform and the respective interface, will be presented in an upcoming deliverable.

Implementation description and status

The implementation of the logging mechanism has been concluded and details about it can be found in the previous version of this deliverable (D5.2). For the auditing mechanism, two different use cases were defined, and three different personas are involved, as described above. To summarize, only the administrator of the platform can check the activities of all users, while organization administrators can check the activities of users belonging to this organization. Logs can be checked based on the data used for each action, only by the organization administrators and the healthcare professionals, under this organization.

Regarding the auditing of the logs, specific functions were implemented in the Smart Contracts and they are responsible of checking whether the user that is executing an audit service, has the necessary rights. Once a user has logged in the platform and has been verified by the blockchain network as known and registered entity, the respective Smart

Contract is invoked. The identity of the user is retrieved from the blockchain (through the Smart Contract) and the user's role is checked against the role that had been specified for this identity at the moment of the user's registration. If the audit service the user tried to execute is allowed for a user with such a role, then control is passed back at the application level and then the query against the MongoDB is executed. If the query is successful, the logs that satisfy the query parameters are returned and presented to the user in the UI. User can optionally specify a timeframe to see the logs for. This workflow is presented in Figure 5.

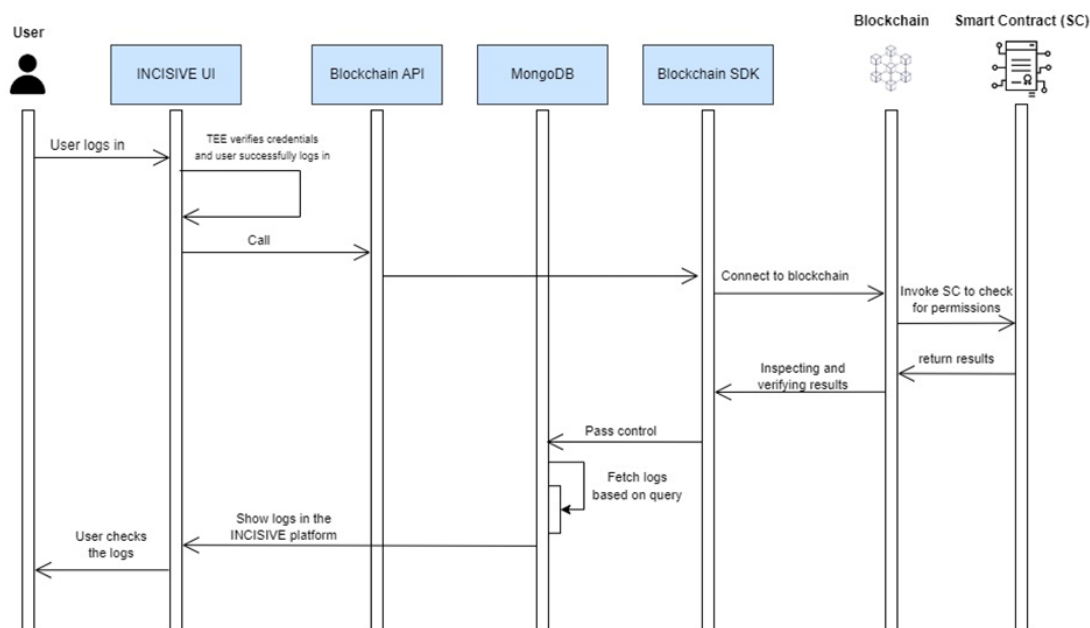


Figure 5. Auditing mechanism workflow.

- **Audit based on user actions**

By default, in case the user is the administrator of the platform, the logs of the actions of all users are presented, and if the timeframe is not specified, from the moment that the component was integrated up to the moment the audit was performed. However, the administrator can search for the actions of a specific user, by specifying the user's username and calling the same service. The same applies to the case of the organization administrator with the difference that the default behavior is that the logs presented are from users that are related to the organization of the administrator. If the organization administrator requests the logs of user that isn't a member of this organization, the response will be empty, as they are forbidden to do so.

- **Audit based on data used**

This service is addressed to users with the role of organization administrators and healthcare professionals. These users, by default, are presented with all the logs that represent actions that include usage of their data. They also have the opportunity to narrow them down by

specifying which data they are interested in seeing how it was used. If they attempt to check actions for data that doesn't belong to their organization, they will receive an empty response. The choice of selecting the dates between which they would like to see the logs for, applies here as well.

For the implementation of the AI models' reputation system, a new Smart Contract has been developed, while exploiting the existing blockchain network. Each of the offered AI models is registered on the blockchain and an initial reputation score of zero points is assigned to them. After using an AI service, medical professionals are prompted to answer a set of questions and their answers are in the Likert scale [11] and act as scores for each question, with the value of "1" representing the lowest value and "5" representing the highest. Each question has a weight assigned, based on the importance of this question. Both the questions and the weights have been defined with the help of AI developers and medical professionals, with weights accumulating to the value of 1. The scores for each question are multiplied by its respective weight, and they are added to a single value. The new reputation score is expressed as the mean value of the old and the new score. This whole procedure is handled by the Smart Contract, ensuring security and transparency, preventing any malicious user of tampering with the reputation of an AI model. The necessary services for this system are also exposed via the API of the Transaction Tracker. The workflow of the reputation system is depicted in Figure 6.

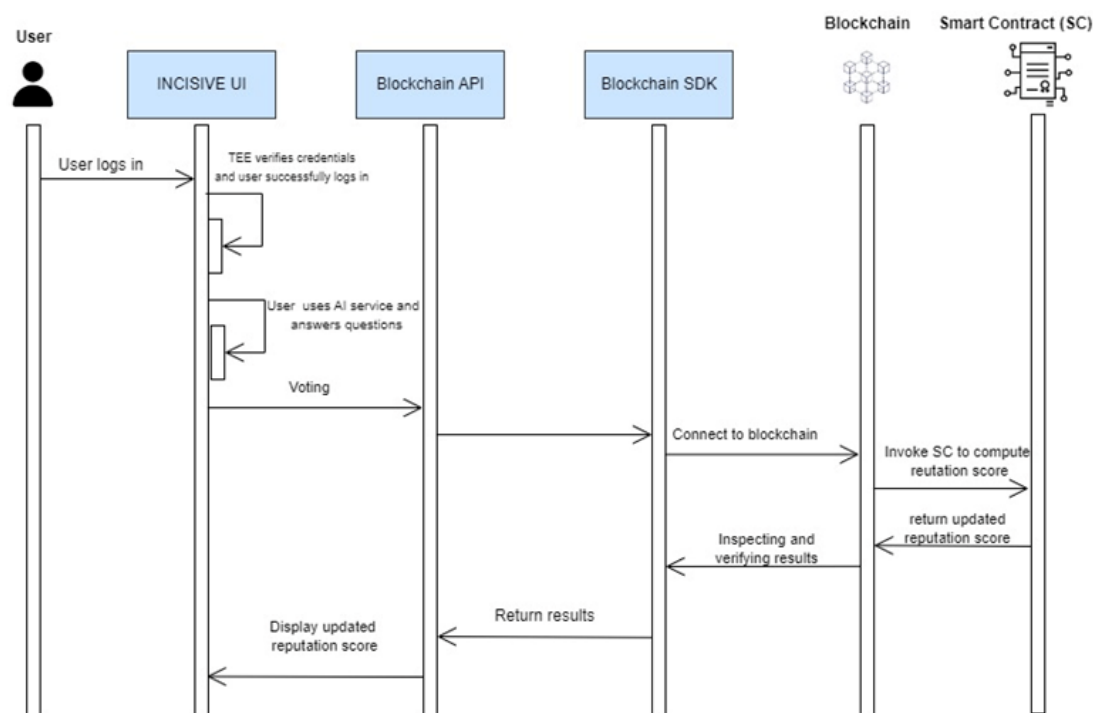


Figure 6. Reputation system workflow.

User needs considered

For the implementation of the Transaction Tracker, the following user needs were considered:

FR18-Availability of a history logging feature to keep track of any changes on the dataset, times of use, downloads, and research conducted on the dataset.

FR-12 Linking the datasets with metadata or explainable tags

BP-IN-10 Security aspects: The system should guarantee the secure management of data, access and transparency of its transactions

BP-IN-12 Results can be traced back to the raw data

Evaluation procedure before integration in the whole platform

For the logging and auditing mechanisms, a testing infrastructure has been set up and their functionalities have been tested by some members and developers of the consortium with dummy data. This procedure has been completed and the component in question has been integrated with the platform and the instances of the ETL tool. The same procedure has been followed for the reputation system, nevertheless it is expected to be continued when the integration with the platform and the creation of the UI screens is completed.

Intended procedure for users' validation

The validation for logging and auditing has taken place. For the reputation system, healthcare professionals will be prompted to test this system by interacting with the UI, answering a set of questions based on a Likert scale (1-5), and thus evaluate the provided functionalities.

4.3. Data sharing schema

4.3.1. Overall description

The data sharing mechanism enables on one hand new data providers (DP) to share their data within the INCISIVE repository and on the other data users (DU) to search and reuse data of the INCISIVE repository. Towards this direction, the data sharing mechanism provides all necessary features, in order to: (1) allow DPs and DUs to become registered users of the INCISIVE repository, (2) align the data sharing process with legal and ethical norms, including data preparation guidelines and support before data is shared, such as data selection, deidentification, curation, etc. (3) provide a data sharing infrastructure, covering all technical and legal aspects of health data sharing, (4) enable data users to search and reuse available data respecting the data access rights, and (5) ensure the DPs the right of opting out. Two main services (use cases) are foreseen for the INCISIVE data sharing users:

1. For Data Providers (DP): Share their data.
2. For Data Users (DU): Search existing data and reuse it.

To support potential INCISIVE data sharing users a dedicated portal has been developed so as to share important information to interested parties. Specifically for the DPs, details on the following aspects are available: (1) the data sharing process; (2) the data access rights; (3) the infrastructure (central or federated node); (4) the data preparation; and (5) the available open tools. Respectively for the DUs: (1) the data use; (2) the data catalogue (including available metadata); (3) criteria on performing a data search; (4) the obligations in terms of acknowledgement (citations) and (5) the conformity to the terms of use.

In order to guarantee the eligibility of potential INCISIVE data sharing users, a registration process has been defined which consists of the following steps:

1. The INCISIVE data sharing user fills in a contact form to express interest in becoming member of the INCISIVE repository as DP or DU.
2. An offline negotiation process will start between the user and a dedicated group (or committee) on behalf of INCISIVE in order to verify the eligibility of the user and finalize their access to the INCISIVE repository.

The data that are shared through INCISIVE can be:

1. **Reusable:** the DUs -once registered- have access to all data shared within INCISIVE.
2. **Private:** some DPs can indicate they have private data to share. This information is available in the data sharing portal. If a DU wants to get access to this data, they have to contact the DP outside the INCISIVE platform and follow an offline process (not technically supported by INCISIVE).

4.3.2. Updates compared with the second prototype

During the second version of this deliverable (D5.2) the design phase was completed, and the data sharing schema was implemented as part of the second INCISIVE prototype (M30). During this period, the data sharing portal has been developed and populated with the appropriate text, internally proof-read and reviewed by the INCISIVE consortium in terms of functionality and information available and deployed on a public domain (<https://share.incisive-project.eu/>).

4.3.3. Functionality implementation according to the foreseen use cases

Data Sharing Mechanism

UC Code	Ref. System	Title
---------	-------------	-------

UC-FS-01	Federated Storage (FS)	Data Sharing
UC-ADM-01	Administration	User Registration

Prototype functionality related to the UC

The DP can initiate a data sharing process, as presented in Figure 7. Towards this direction and for the better support of the DP, guidelines are available related with the data preparation and uploading procedures. Indicatively, this information includes: (1) information about what a DP gains when sharing data (incentives); (2) Information on the specifications of the data (metadata); (3) Information about the data preparation steps; (4) Information about the available access levels supported by the platform; (5) Information on the de-identification protocol; (6) Link to INCISIVE de-identification tool; (7) Information on semi-automatic annotation ; (8) Link to INCISIVE annotation tool; (9) Information on data quality checking process; (10) Guidelines to fill-in and upload the data templates; (11) Link to download the data templates; and (12) INCISIVE support provision (either on-line or off-line).

This process may be repeated as many times as the DP has new data to share with INCISIVE.

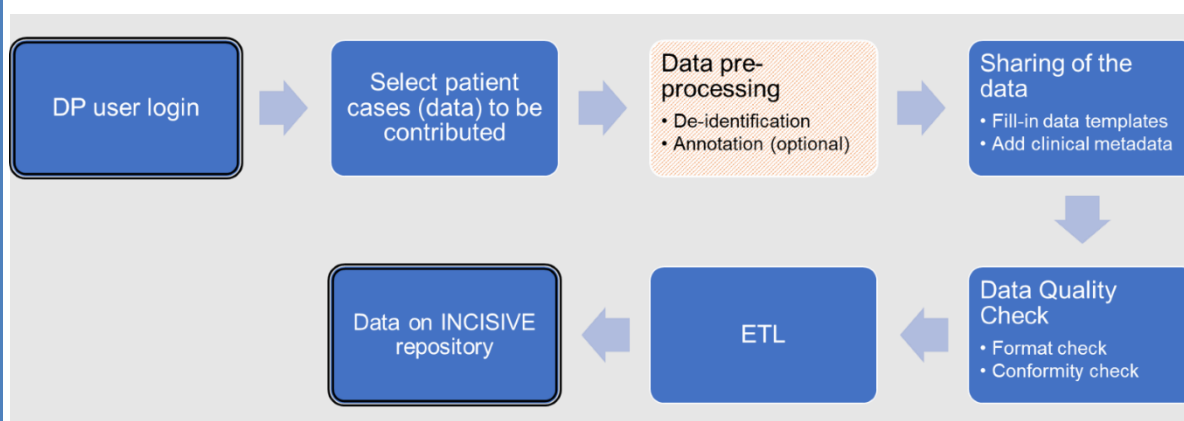


Figure 7. Data sharing process.

Technical description

The UML diagram below (Figure 8) presents how the data sharing mechanism works:

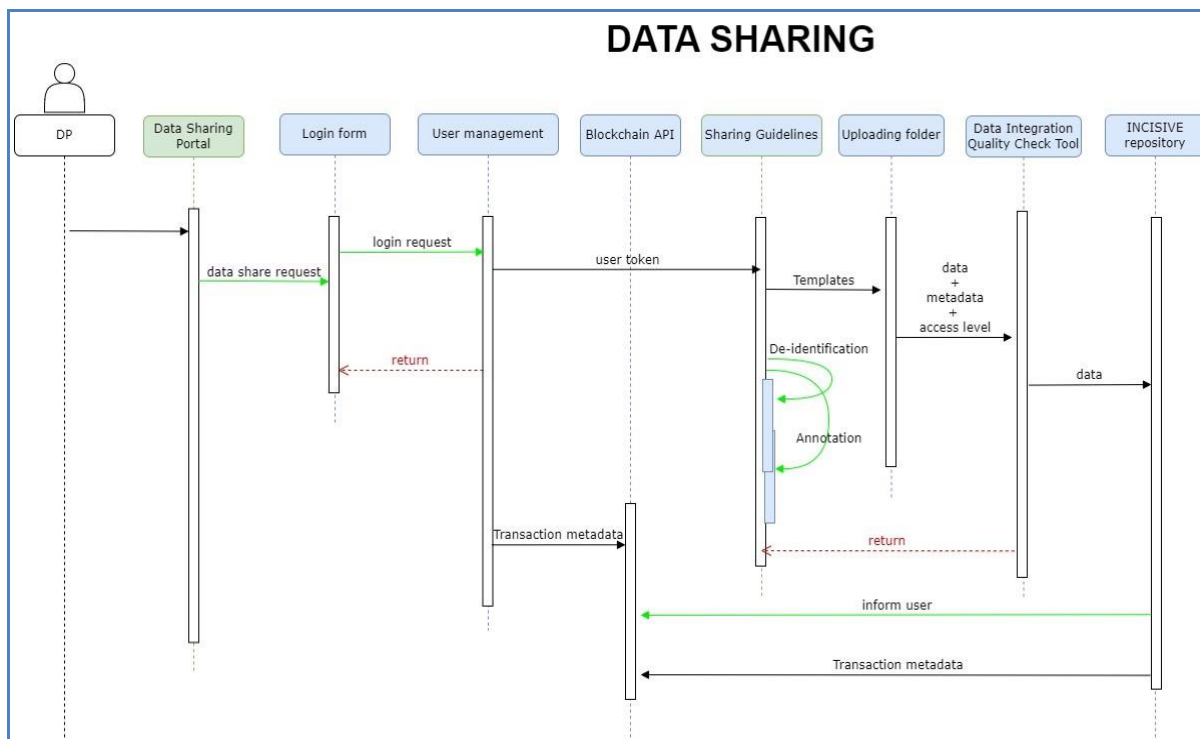


Figure 8. Data sharing mechanism.

The DP visits the data sharing portal and logs in through the corresponding form provided by the platform. The system checks whether the user has the rights to access the mechanism. If this is correct, the user is directed to the INCISIVE Repository UI and the login transaction is written to the blockchain. The user (DP) now has access to the functionalities for sharing data within the INCISIVE platform. These include data sharing templates and guidelines, de-identification and annotation tools. Before being able to consume these services, the user must have accepted the terms and conditions of the platform. This is performed during the registration procedure (Figure 9). Upon completion of the templates and the performance of annotation and de-identification processes, the quality check of the data takes place. If this is successful, the data is uploaded to the platform, while the transaction is recorded on the blockchain, and the user is notified of the successful completion of the sharing process. Otherwise, the control returns back to the uploading point, where the user needs to perform further modifications on the data to successfully complete this process.

Dependencies / interfaces

A user registration process should have taken place before the user is able to use the data sharing mechanism (as described in section 4.3.1). As already described in D2.5, the services will be provided through a specific interface. Before any action takes place, the data provider should confirm that he/she is informed about the overall data sharing framework, and the actions required before registering into the platform. Following this, it is assumed

that offline preparatory actions between the user and the platform took place before the initiation of the data sharing process.

Two types of data providers (DP) are foreseen, namely individuals and organizations. In case of Individual DP, we should expect that they will connect to the central repository and not install a Federated Node. Organizations will be able to select either of the cases we support in the hybrid scheme (access through central or federated node).

The DP registration journey is presented in Figure 9.

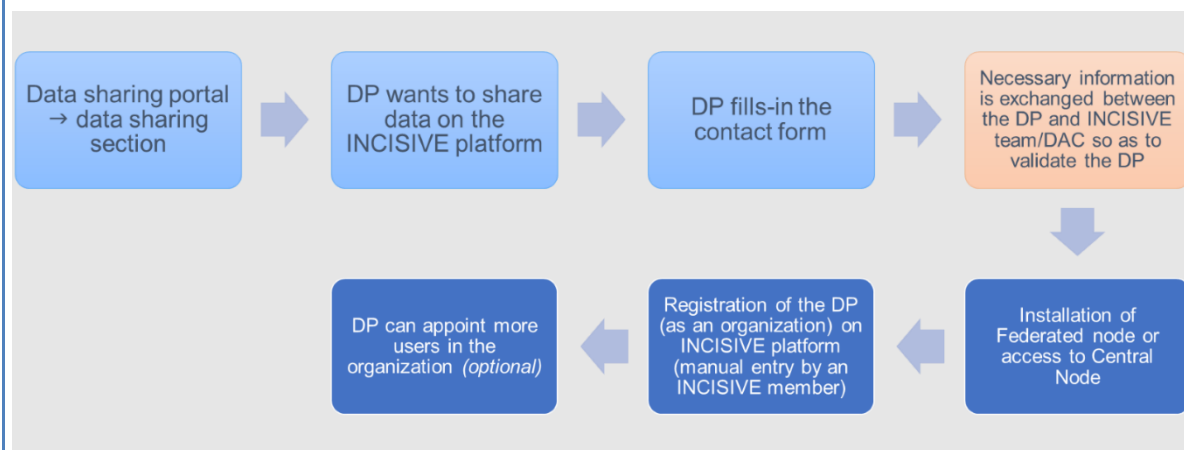


Figure 9. Data Provider's registration process.

Implementation description and status

The data sharing portal has been implemented, proof-read and reviewed (in terms of functionality and information available) by the INCISIVE consortium. The portal is available on <https://share.incisive-project.eu/>.

User needs considered

With this functionality, the following user needs are considered (business processes (BP – D5.2):

- BP-IN-09 Legal and ethical aspects
- BP-IN-10 Security aspects: The system should guarantee the secure management of data, access and transparency of its transactions
- BP-IN-13 Existing sources can be linked with the INCISIVE platform
- BP-IN-14 Data should be de-identified before any process
- BP-IN-16 Data sharing
- BP-IN-25 Annotation of medical images
- BP-IN-31 Data quality check

Evaluation procedure before integration in the whole platform

The data sharing mechanism has been tested by the DPs and is ready for use. The data sharing portal has been reviewed by the INCISIVE consortium members and is ready for use.

Intended procedure for users' validation

Open ended feedback has been collected by the INCISIVE consortium members regarding the data sharing portal. Text review has been also performed by members of the INCISIVE consortium.

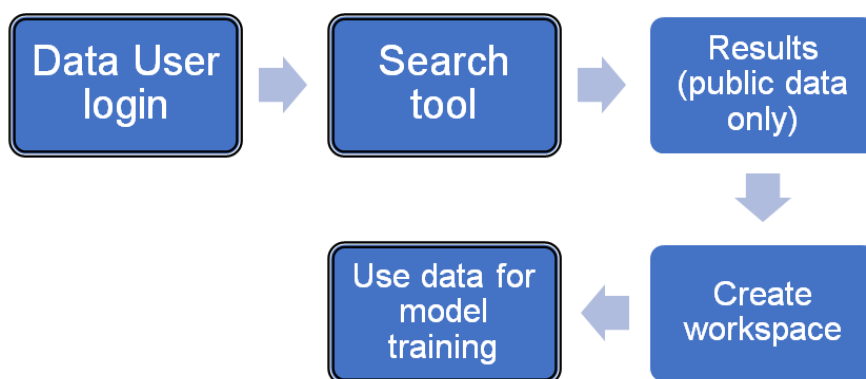
The evaluation of the data sharing mechanism will be part of the evaluation of the INCISIVE integrated final prototype (D6.3) due M36 based on the related evaluation protocol already defined under WP6.

Data Use Mechanism

UC Code	Ref. System	Title
UC-FS-01	Federated Storage (FS)	Data Sharing
UC-FS-02	Federated Storage (FS)	Collection of Data upon a specific query /request
UC-ADM-01	Administration	User Registration

Prototype functionality related to the UC

The data users (DU) may search for data that are on the INCISIVE platform using the search functionality (as already presented). S/he submits a search query and gets back results that meet her/his criteria. The DU may view the data before s/he selects it for use. Every time a DU uses data from another DP (DP source), they will have to acknowledge the DP.



INCISIVE platform
 Actions tracked by blockchain

Figure 10. Data Use process.

Technical description

The data use process is depicted in the following UML diagram (Figure 11).

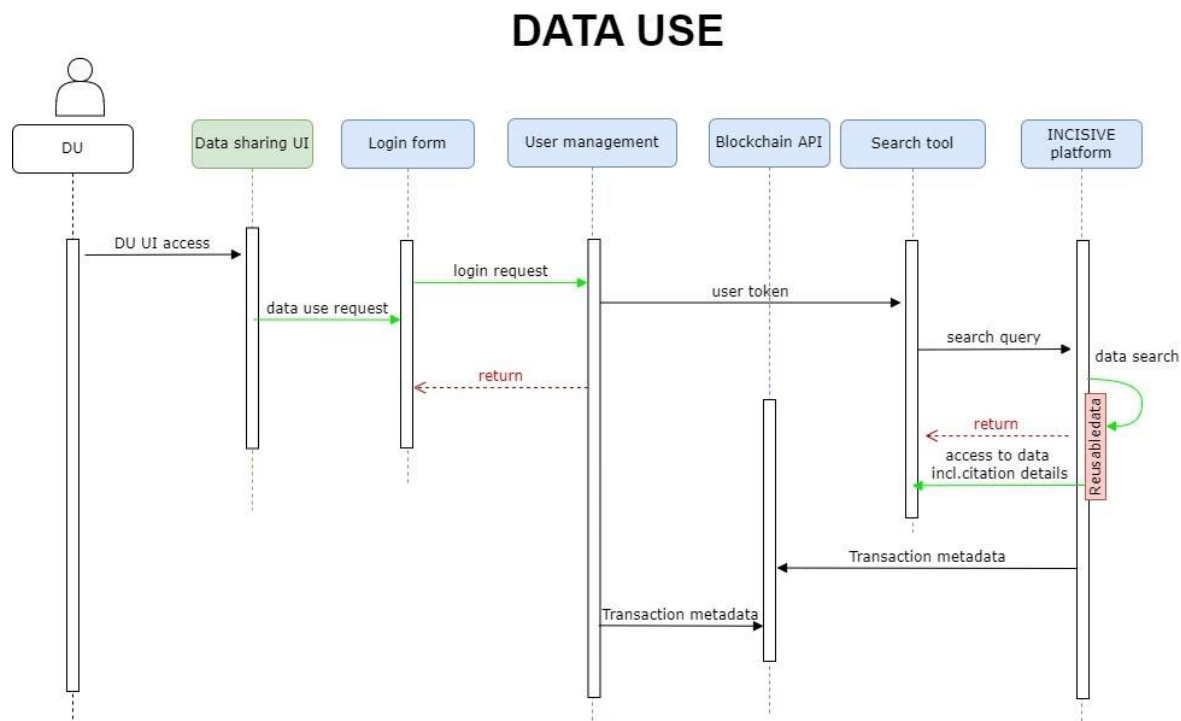


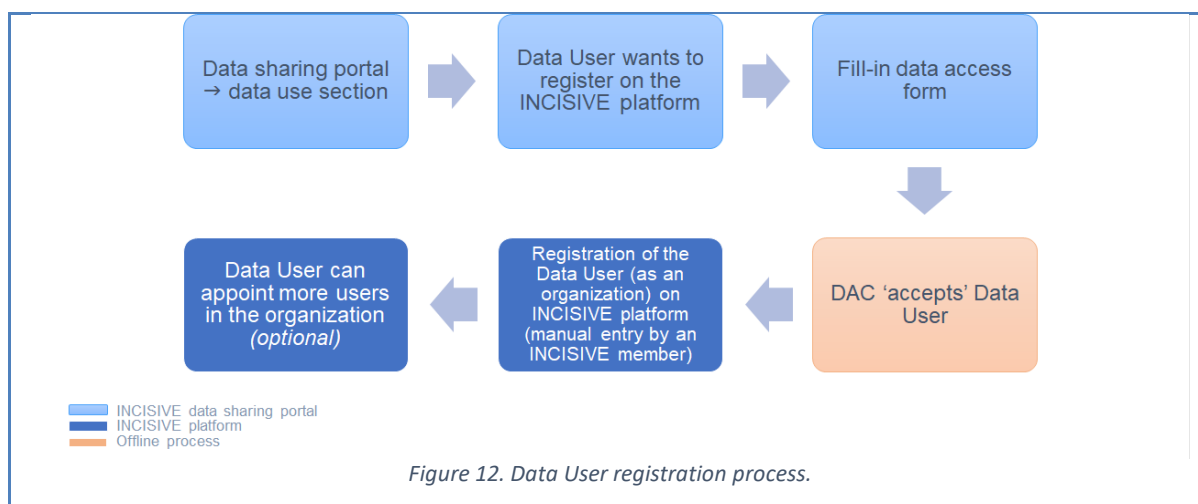
Figure 11. Data Use mechanism.

The login functionality for the DU is similar to the one presented in the data sharing use case. In the corresponding UI the user can perform a data use request by submitting a search query to the INCISIVE platform for reusable data (available in the platform) that fulfil certain search criteria. In case that such data exists, access to data is provided combined with citation details. All these transactions are logged to the blockchain for auditing purposes.

Dependencies / interfaces

A user registration process should have taken place before the user to be able to use the data sharing mechanism (as described in section 4.3.1). As already described in D2.5, the services will be provided through a specific interface.

The DU registration journey is presented in Figure 12.



Implementation description and status

The data sharing portal has been implemented, proof read and reviewed (in terms of functionality and information available) by the INCISIVE consortium. The portal is available on <https://share.incisive-project.eu/>.

User needs considered

With this functionality, the following user needs are considered (business processes (BP – D5.2):

BP-IN-09 Legal and ethical aspects

BP-IN-10 Security aspects: The system should guarantee the secure management of data, access and transparency of its transactions

BP-IN-13 Existing sources can be linked with the INCISIVE platform

BP-IN-16 Data sharing

BP-IN-33 Dataset documentation and description

Evaluation procedure before integration in the whole platform

The data sharing mechanism has been tested by the DUs and is ready for use.

The data sharing portal has been reviewed by the INCISIVE consortium members and is ready for use.

Intended procedure for users' validation

Open ended feedback has been collected by the INCISIVE consortium members regarding the data sharing portal. Text review has been also performed by members of the INCISIVE consortium.

The evaluation of the data sharing mechanism will be part of the evaluation of the INCISIVE integrated final prototype (D6.3) due M36 based on the related evaluation protocol already defined under WP6.

5. INCISIVE Repository user interface

In this chapter the updated UIs of the INCISIVE platform and the DIQCT are presented. For the sake of the document’s completeness, the UIs of the data de-identification and annotation tools are also included.

5.1. Data preparation: quality check, de-identification and annotation

5.1.1. DICOM De-identification and curation tool

The data de-identification tool has its own interface and a simple, user-friendly UI, shown in Figure 13. It is accessible through this URL: <https://dicom-de-identification-and-curation-tool.incisive.iti.gr/>. More information about the de-identification tool can be found in D5.2.



Figure 13. De-identification and curation tool interface.

5.1.2. Semi-automatic annotation tool

The semi-automatic annotation tool is a standalone tool with its own interface, shown in Figure 14. Users can access it by following this link: <https://semi-automatic-annotation-tool.incisive.iti.gr/>. More information about the semi-automatic annotation tool can be found in D5.2.

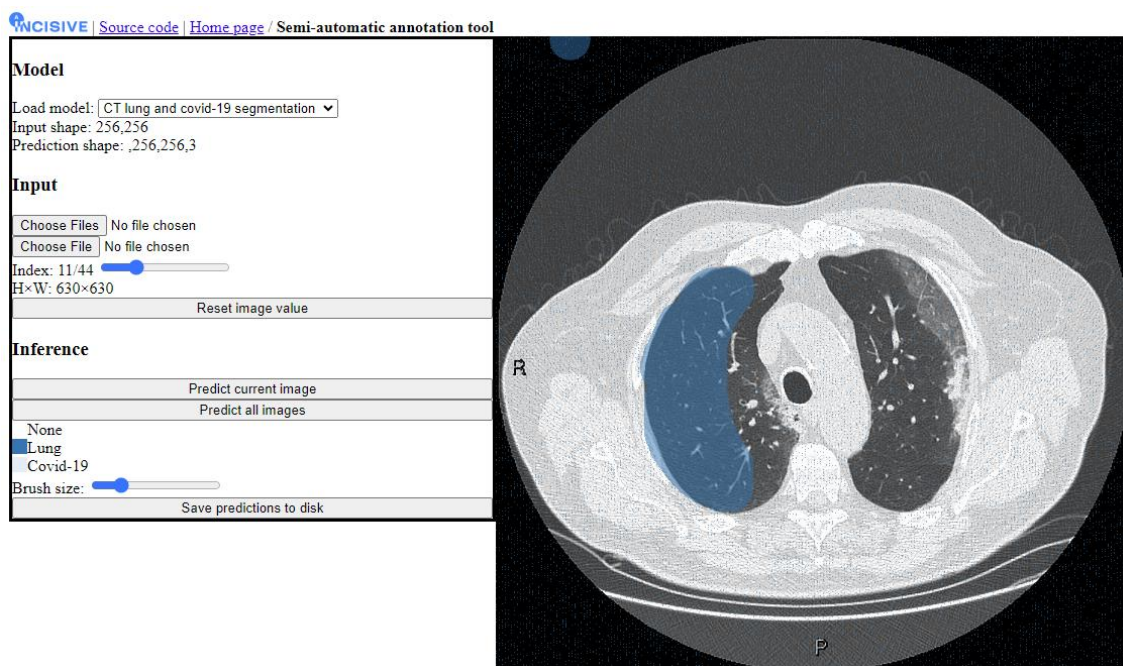


Figure 14. Semi-automatic annotation tool interface The blue area indicates the region of the lung (in this case), that has been annotated, using the brush tool.

5.1.3. Data Integration Quality Check Tool

Collected data are stored in the local premises of the user that intends to upload them into the INCISIVE platform. In order to integrate the data into the INCISIVE platform, a data quality check must be applied. The goal of the check is to identify whether the data follow the harmonization requirements reflecting, among others, the de-identification protocol applied, the inclusion of all the required information on the accompanying metadata as well as the integrity. The folder structure is also checked. A consistency check between the case metadata in the imaging (DICOM) patient data also take place. The outcome of the data quality check are reports proposing corrective actions to the user.

Pipeline followed by end-user

1. The user must login to the application.
2. The user must select the data needed to be analysed from the parent folder where the data lies, the images and the excel file.

3. The user must check the data in terms of **Clinical Metadata Integrity**. This process reveals errors related to the template structure (Tabs and Columns) and Patient codification (Proper patient encoding and duplicate patient ids), as well as the content of the template, meaning the values inserted in each field. The user needs to go back to the template and correct the errors reported by the tool.
4. The user must check the data in terms of **Case Completeness**. This component presents an overview of the data provided, what modalities in each timepoint, as well as the percentage of mandatory fields that are present for each patient. No action is required by the user in this stage, but this step is needed to be executed because it is a dependency for the next step.
5. The user must check the data in terms of **Template-Image Consistency** i.e., if the file structure of the images is compliant to the template. This component performs a proper renaming of the studies' folders so they can be stored in a unified and commonly understandable way, which is a prerequisite in INCISIVE. The user needs to go back to the data and correct the errors reported by the tool.
6. The user must check the data in terms of **DICOM de-identification Protocol**. This component checks whether the de-identification protocol, already defined in INCISIVE, is properly applied in the metadata included on the imaging files. If errors appear in this step, the user must communicate with the respective partner and resolve the issue.
7. The user must check the data in terms of **DICOM validation, i.e., if the provided DICOM images are valid DICOM files**.
8. The user must check the data in terms of **DICOM Analysis Requirements** i.e., expected quality for analysis and the training of the algorithms and the proper placing of the annotation file inside the correct series folder. The user needs to go back to the data and correct the errors reported by the tool.
9. The user must check the data in terms of **DICOM Overall Patient evaluation**. This component checks if there are duplicate images in the dataset and if some prerequisites for the analysis are present, modalities in each cancer type, slice thickness and pulse sequence. The user needs to go back to the data and correct the errors reported by the tool.
10. The user needs to run the pipeline again to ensure that all errors are corrected.

Steps 1, 7 & 8 are newly added components in relation to the previous version and Figure 15, Figure 16 and Figure 17 below depict their implementation in the user interface.

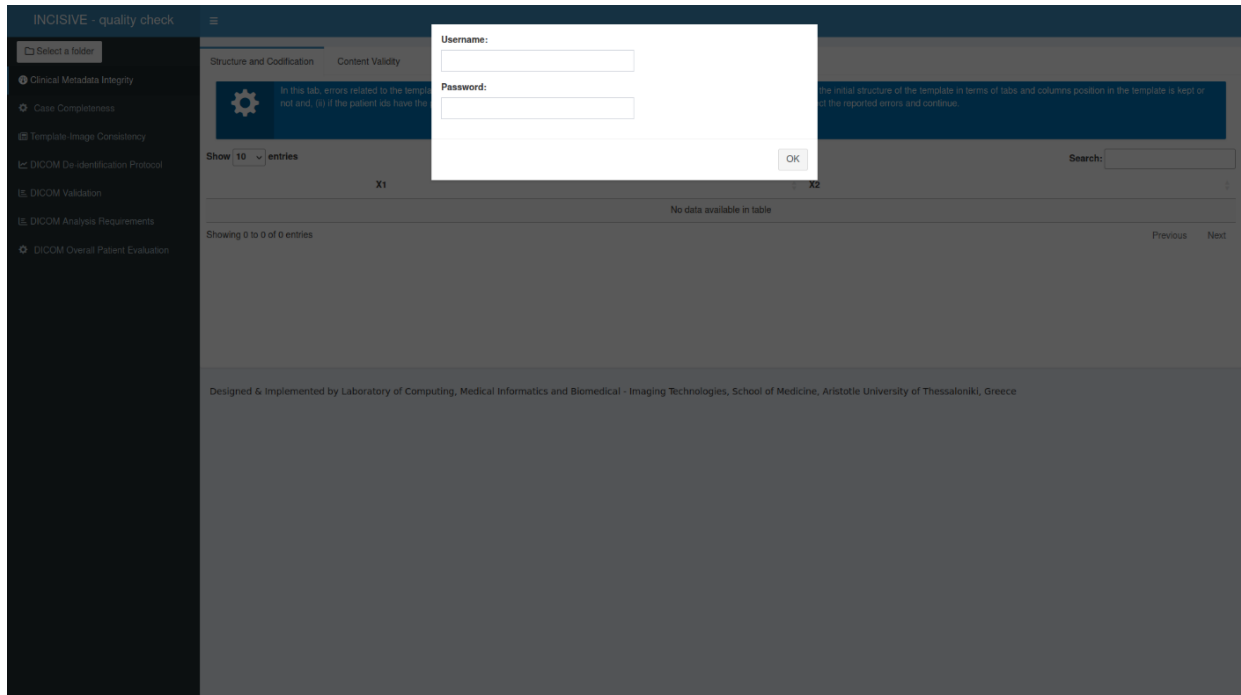


Figure 15. The login page of the DIQCT.

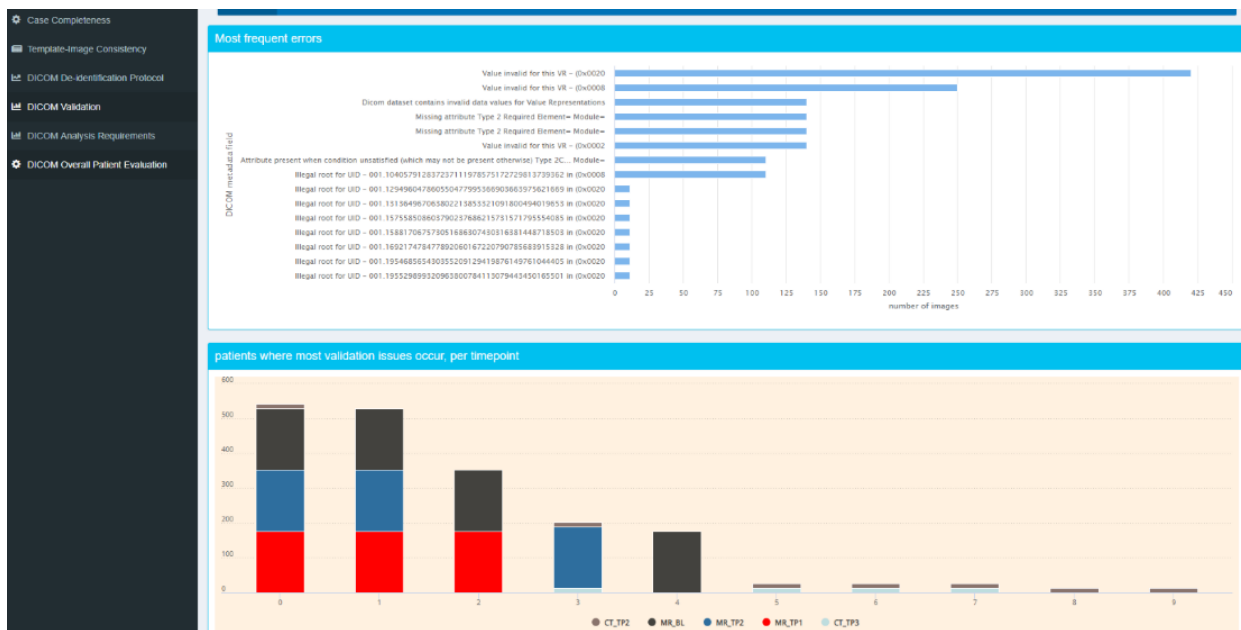


Figure 16. The results of the newly added DICOM Validation Component.

The screenshot shows the 'INCISIVE - quality check' interface. On the left is a navigation menu with options like 'Clinical Metadata Integrity', 'Case Completeness', etc. The main area is titled 'Annotations' and contains a table of results. A message at the top states: 'This component checks if there are annotations in images'. The table has columns for 'patientID', 'timepoint', 'series', and 'filename'. Five rows of data are shown, with the first two highlighted in orange and the third in red. The bottom of the table indicates 'Showing 1 to 5 of 5 entries' and includes 'Previous' and 'Next' navigation buttons.

	patientID	timepoint	series	filename
1	001-000224	001-000224_CT_BL	Series-1	001-00001-BL_PETCT.nii.gz
2	001-000224	001-000224_CT_BL	Series-1	Untitled.nii.gz
3	001-000224	001-000224_CT_TP1	Series-1	Untitled.nii.gz
4	001-000224	001-000224_CT_TP2	Series-1	Untitled.nii.gz
5	001-000224	001-000224_CT_TP3	Series-1	Untitled.nii.gz

Figure 17. The results of the newly added Annotation Component.

5.2. INCISIVE data search

The user, once logged in to the INCISIVE portal, can utilize the available functionalities, according to their role. One of the core functionalities, involved searching for the available data in the platform via the Search tab on the menu. By selecting the “Search” option, the user is redirected to the corresponding page where they need to fill in information related to the data they would like to search for, as shown in Figure 18.

The screenshot shows the 'INCISIVE' search page. It features a navigation menu on the left with 'Search' selected. The main content area is titled 'Search' and includes a description: 'The search functionality allows you to query the data available in INCISIVE. Each search being performed is distributed among data providers and the results are aggregated by INCISIVE.' Below this is a 'Search' button. The form is divided into two sections: 'Patient Information' and 'Cancer Information'. The 'Patient Information' section has fields for 'Gender' (a dropdown menu), 'Age From', and 'Age To'. The 'Cancer Information' section has fields for 'Cancer Type' (a dropdown menu), 'Cancer Stage' (a dropdown menu), and 'Months of Observation'.

Figure 18. Performing a search in the platform.

The search includes three main categories of information the user may define: (i) patient information, such as gender, and age, (ii) cancer information such as cancer type, stage of cancer which is relative to the cancer type, image modalities, etc., and finally (iii) data

information including to which data providers should the search be performed, country of origin, and whether genomic data are available or not.

Once the search is performed, the user is informed that once the search is over, they may proceed to the “Search results” page in order to view a comprehensive list of the currently ongoing searches as well as the finished ones, as depicted in Figure 19. As also depicted, the ongoing searches are marked with the “RUNNING” status, while the finished searches are marked with the “FINISHED” status.

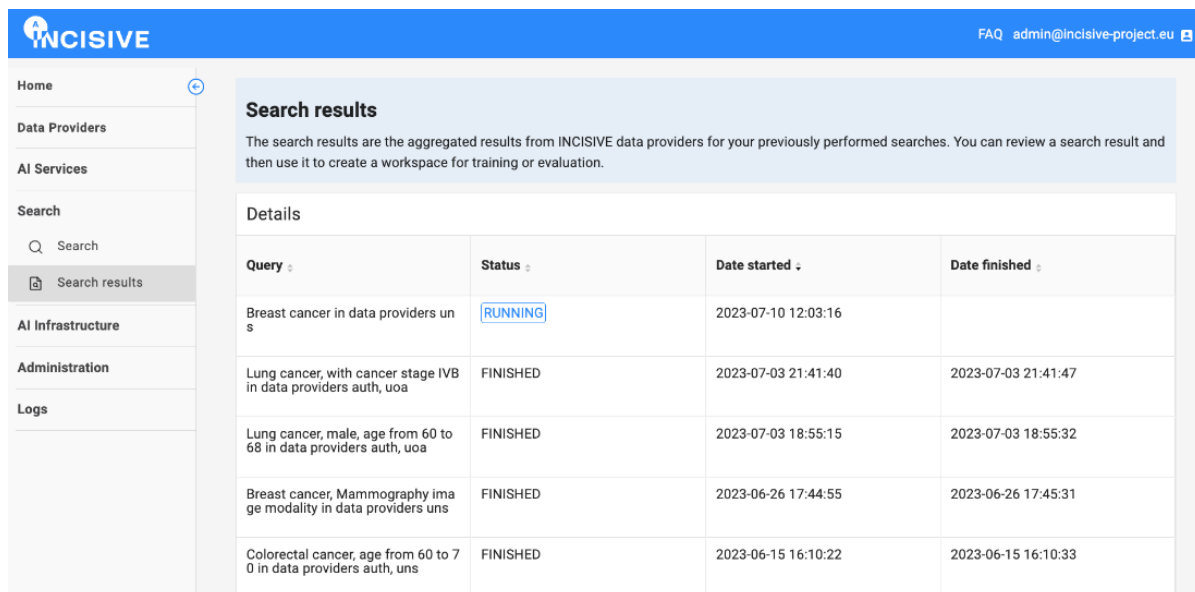
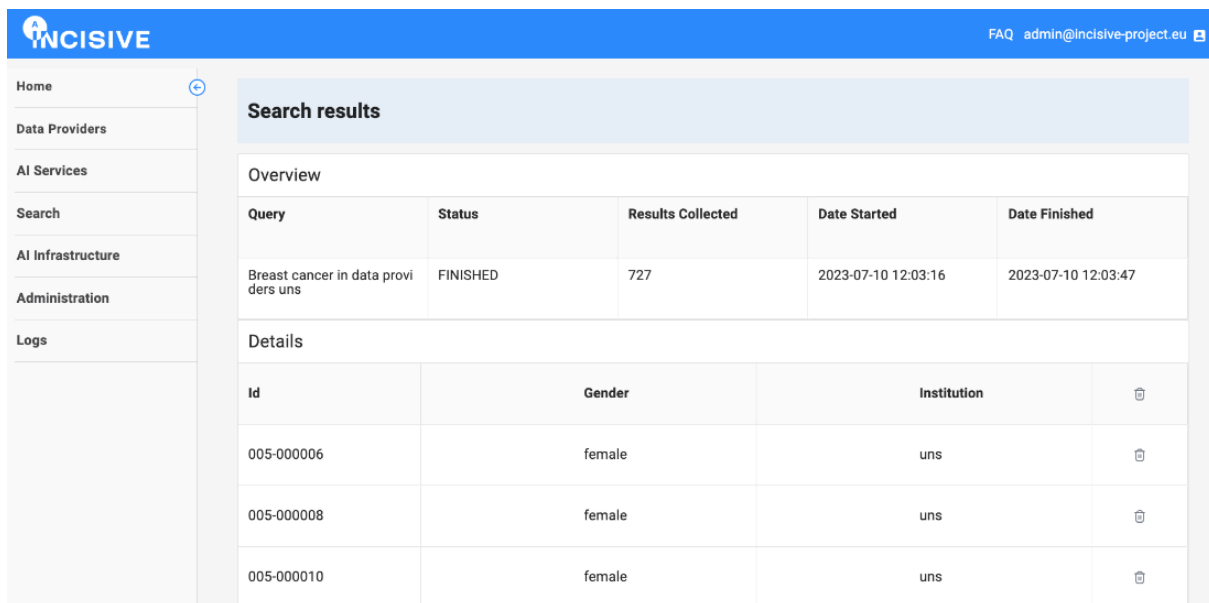


Figure 19. Search results page.

Consequently, the user may select any of the available searches in order to view detailed information related to the results, as it can be seen in Figure 20. Summarized information related to the search can be viewed at the “Overview” section, including number of patients, when was the search performed, and from which data providers the data are originated. Furthermore, for each patient, the user may see more detailed information by clicking on them on the “Details” section of the page, including also thumbnails of the images of a patient.



The screenshot shows the INCISIVE web interface. On the left is a navigation menu with items: Home, Data Providers, AI Services, Search, AI Infrastructure, Administration, and Logs. The main content area is titled 'Search results' and contains an 'Overview' table and a 'Details' table.

Query	Status	Results Collected	Date Started	Date Finished
Breast cancer in data providers uns	FINISHED	727	2023-07-10 12:03:16	2023-07-10 12:03:47




Id	Gender	Institution	
005-000006	female	uns	
005-000008	female	uns	
005-000010	female	uns	

Figure 20. View detailed information of a search.

In addition, the user may filter the produced results using the “Filter Results” button by percentage, or delete cases / patients they may not want by clicking on the corresponding icon.

5.3. INCISIVE workspaces

Once a search is performed and the results are produced, the user may select to create a Workspace. This can be performed by clicking the “Create Workspace” button in the selected search result’s page, as shown in Figure 21:

ID	Gender	Status	Action
005-000010	female	uns	[trash]
005-000020	female	uns	[trash]
005-000028	female	uns	[trash]
005-000031	female	uns	[trash]
005-000037	female	uns	[trash]
005-000851	female	uns	[trash]
005-000852	female	uns	[trash]
005-000857	female	uns	[trash]

Items per page: 10 | 1 - 10 of 727

[CREATE WORKSPACE](#) [FILTER RESULTS](#) [DELETE](#)

Figure 21. Select the creation of a Workspace from a search.

Afterwards, the users should fill in additional details regarding the workspace, including its name and description, which AI engine will be used and for what purpose (i.e., to train a model, to evaluate it or to fine tune one). Based on these options they will need to fill in additional information related to the actual model, such as its AI engine and version, its output, and its configuration. The detailed information the user should fill in can be found in Figure 22:

Create workspace

Data to use

Search	Total Results	Data Providers
Breast cancer in data providers uns	727	1

Workspace Details

Name* Description* Type*

Workspace AI Engine & models

Ai Engine* Ai Engine Version* Output AI Model Name*

The AI Engine to be used in your workspace The AI Engine Version to be used in your workspace The name of the ai model to be created

Figure 22. Create a Workspace.

By clicking on the “Create Workspace” button once this information is filled, the new Workspace is now created and available for the user to utilize. The list of the available Workspaces can be found under the AI Infrastructure category, as also shown in Figure 23, where similar to the Search results section, the user may select of the available options (i.e., workspaces) in order to view additional information for them:

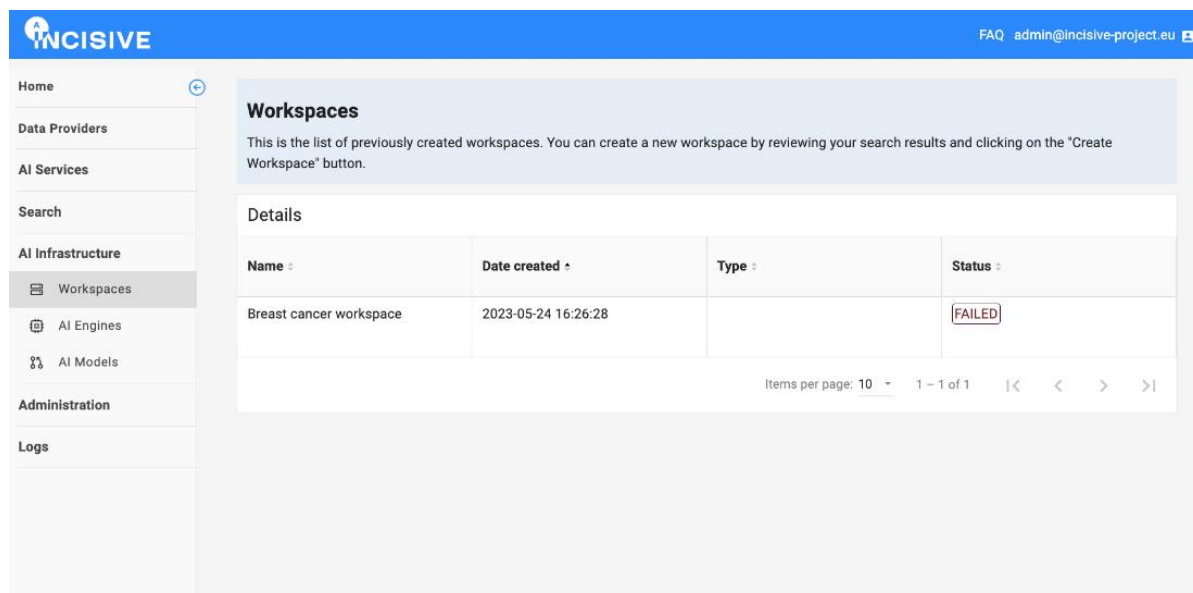
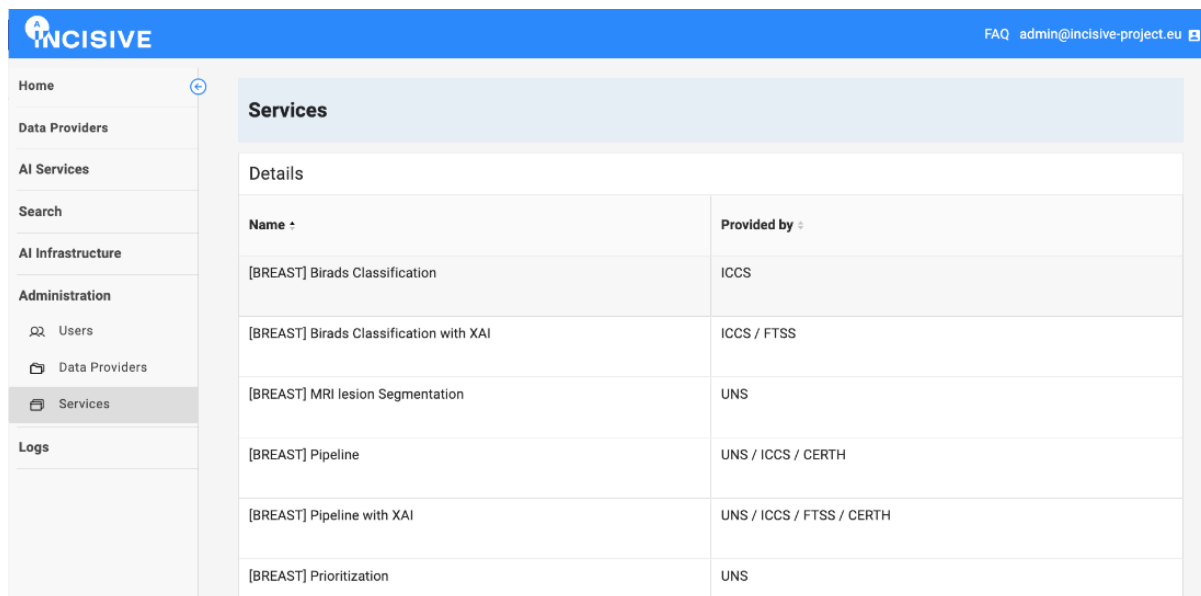


Figure 23. List of available Workspaces.

5.4. Administration

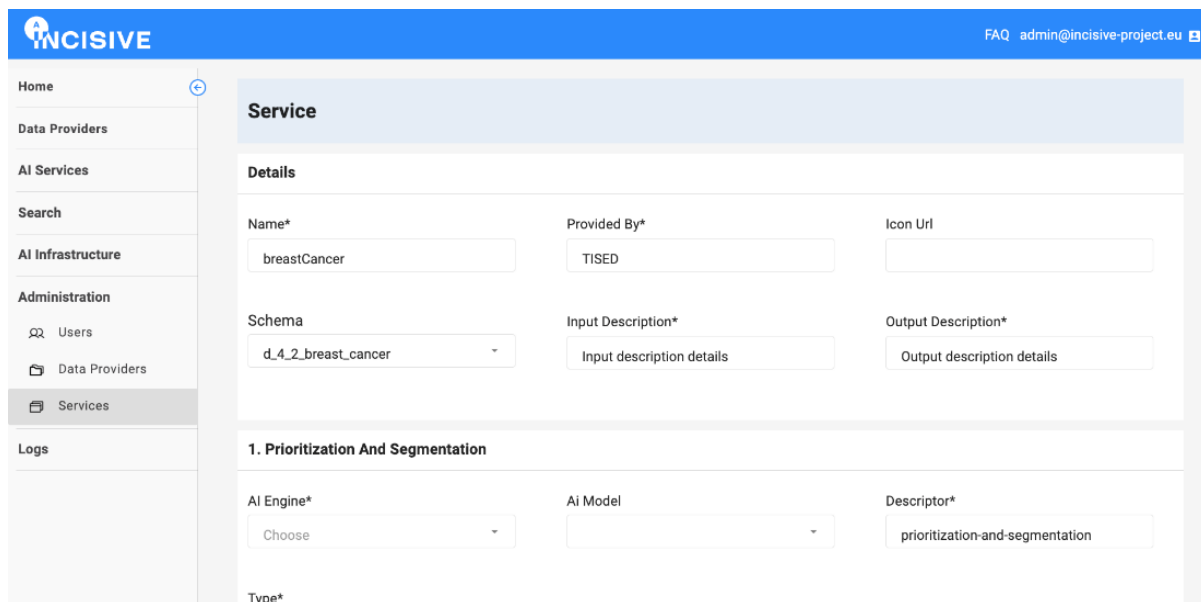
The organization administrators and AI developers/researchers have access to the Administrative section of the platform, allowing them to create new AI Services which can then be utilized by Healthcare Professionals. In the Services page the AI developers may view a list of the available AI Services and select to create new individual Services, or Pipelines, as shown in Figure 24.



Services	
Details	
Name	Provided by
[BREAST] Birads Classification	ICCS
[BREAST] Birads Classification with XAI	ICCS / FTSS
[BREAST] MRI lesion Segmentation	UNS
[BREAST] Pipeline	UNS / ICCS / CERTH
[BREAST] Pipeline with XAI	UNS / ICCS / FTSS / CERTH
[BREAST] Prioritization	UNS

Figure 24. List of available Services.

In order to create a new AI Service, the user will have to fill in information related to it, such as its name, required input / output description, its schema, etc. According to the schema, related information may also be required. For instance, as shown in Figure 25 the user wants to create a breast pipeline, hence 4 steps are created, starting from step 1 which regards a Prioritization and Segmentation service.



Service

Details

Name* Provided By* Icon Url

Schema Input Description* Output Description*

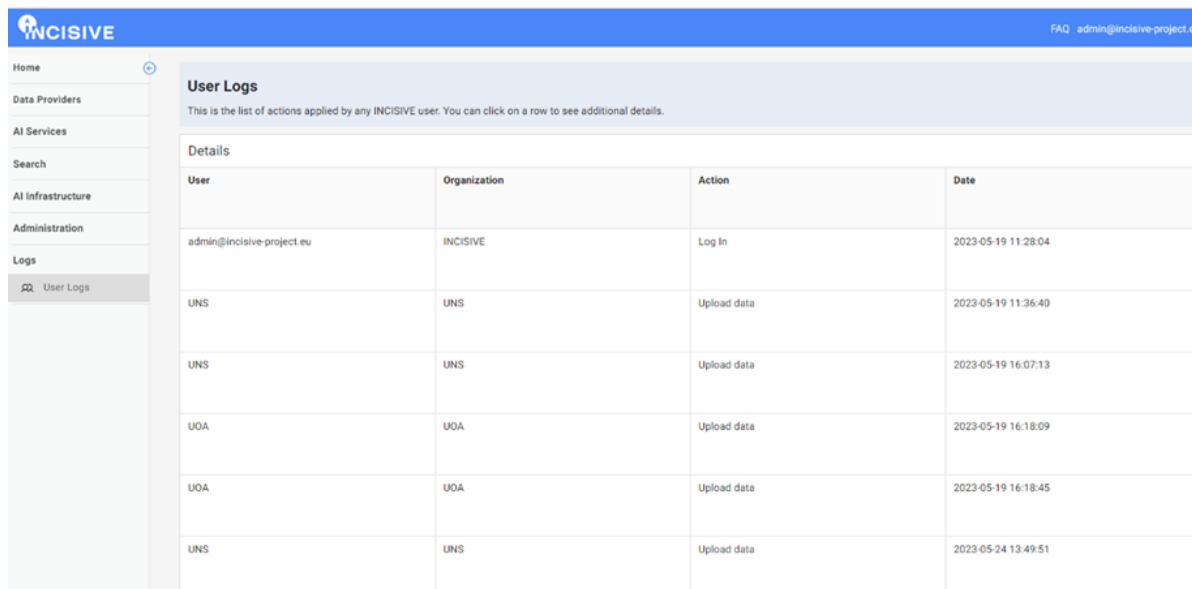
1. Prioritization And Segmentation

AI Engine* AI Model Descriptor*

Type*

Figure 25. Create a new AI Service or a Pipeline.

Users with administrative rights in the platform (administrators of the platform, organization administrators) and healthcare professionals have access to the auditing mechanism, where they can inspect the logs, as shown in Figure 26:



User	Organization	Action	Date
admin@incisive-project.eu	INCISIVE	Log In	2023-05-19 11:28:04
UNS	UNS	Upload data	2023-05-19 11:36:40
UNS	UNS	Upload data	2023-05-19 16:07:13
UOA	UOA	Upload data	2023-05-19 16:18:09
UOA	UOA	Upload data	2023-05-19 16:18:45
UNS	UNS	Upload data	2023-05-24 13:49:51

Figure 26. Logs inspection from the administrator.

5.5. Data sharing portal

The data sharing mechanism is accessible through a web-based specific interface. Despite being an external tool, the UI follows the branding guidelines of the project (e.g., colours, fonts etc), in order to ensure the continuity of the user experience. The design of the Data Sharing portal is responsive, enabling its accessibility from different devices. A screenshot of the data sharing mechanism home page is presented below (Figure 27):

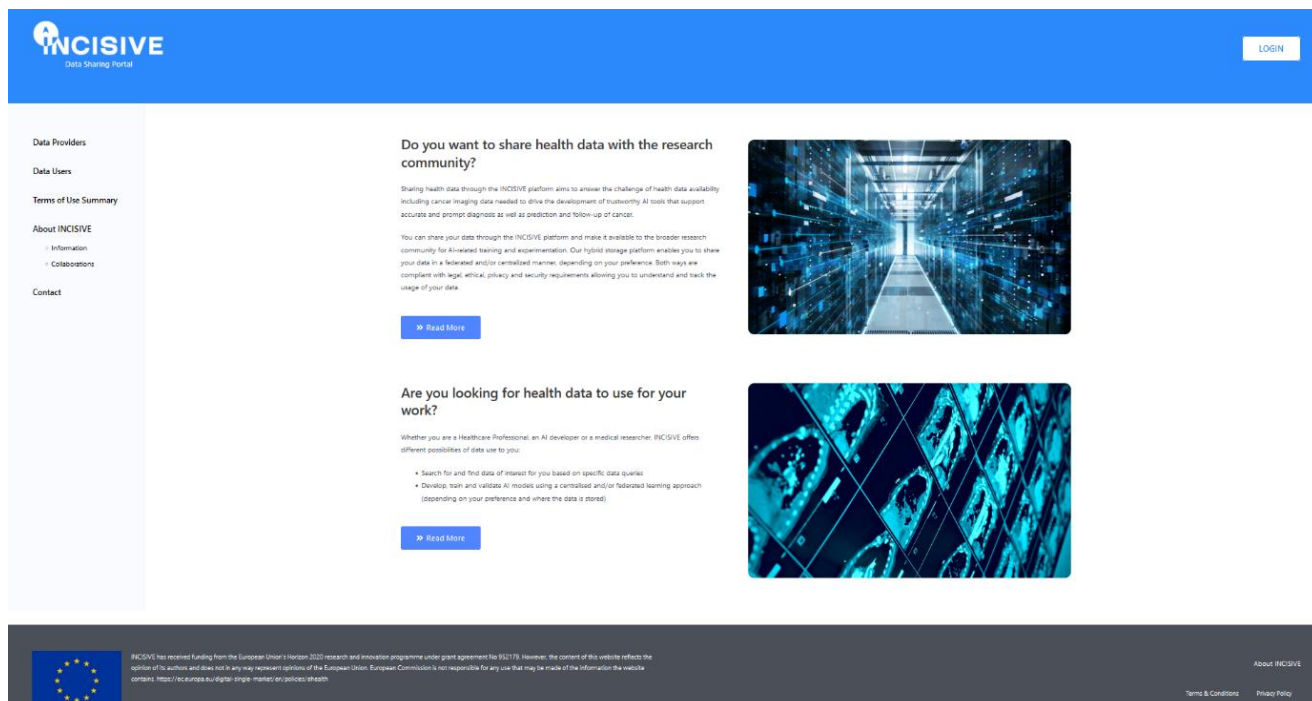
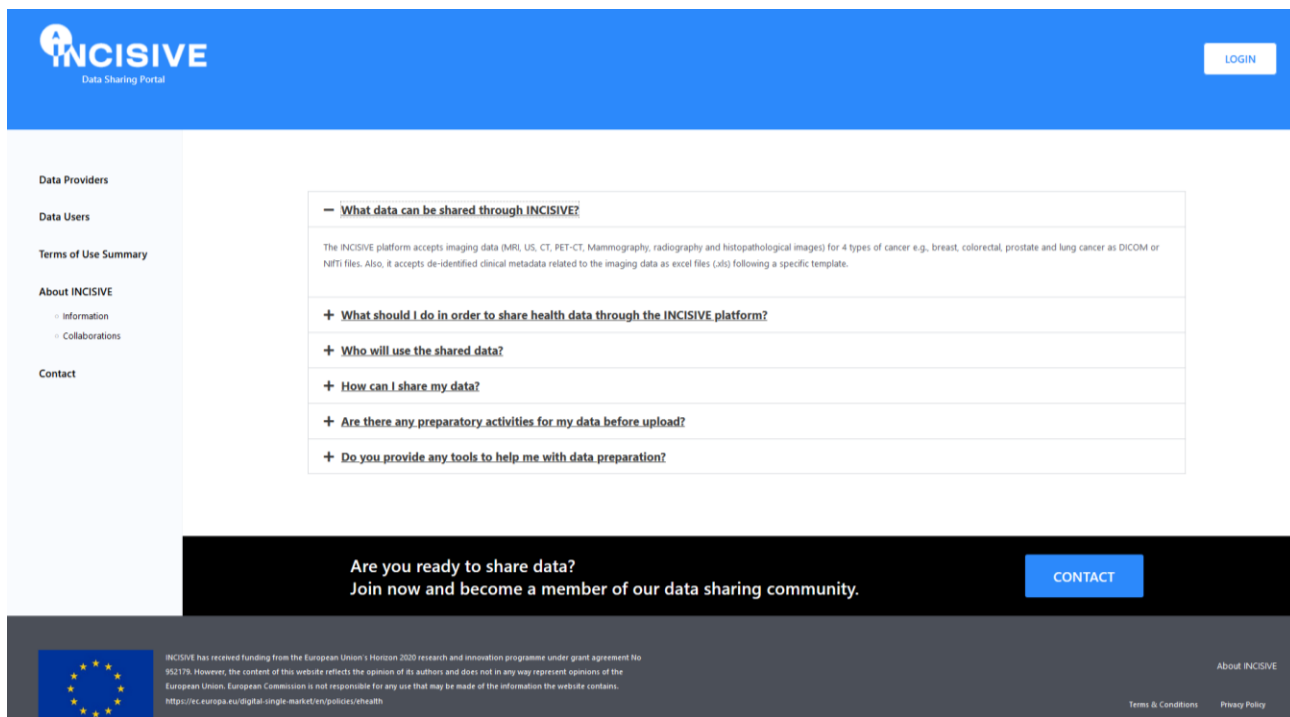


Figure 27. Home-page of the Data Sharing portal.

The UI provides open guidelines and details on the Data Providers and Data Users before making any use of the INCISIVE services, combined with general information on the project, its scope and objectives. This information accessible through the menu on the left part of the screen. More specifically, the available options for the Data Providers (Figure 28) concern: (1) the data sharing process; (2) the data access rights; (3) the infrastructure (central or federated node); (4) the data preparation; and (5) the available open tools. For the Data Users the information available (Figure 29) includes: (1) the data use; (2) the data catalogue (including available metadata); (3) criteria on performing a data search; (4) the obligations in terms of acknowledgement (citations) and (5) the conformity to the terms of use.



INCISIVE
Data Sharing Portal

[LOGIN](#)

- Data Providers
- Data Users
- Terms of Use Summary
- About INCISIVE
 - Information
 - Collaborations
- Contact

What data can be shared through INCISIVE?

The INCISIVE platform accepts imaging data (MRI, US, CT, PET-CT, Mammography, radiography and histopathological images) for 4 types of cancer e.g., breast, colorectal, prostate and lung cancer as DICOM or NIFTI files. Also, it accepts de-identified clinical metadata related to the imaging data as excel files (.xls) following a specific template.

What should I do in order to share health data through the INCISIVE platform?

Who will use the shared data?


How can I share my data?

Are there any preparatory activities for my data before upload?

Do you provide any tools to help me with data preparation?

Are you ready to share data?
 Join now and become a member of our data sharing community.

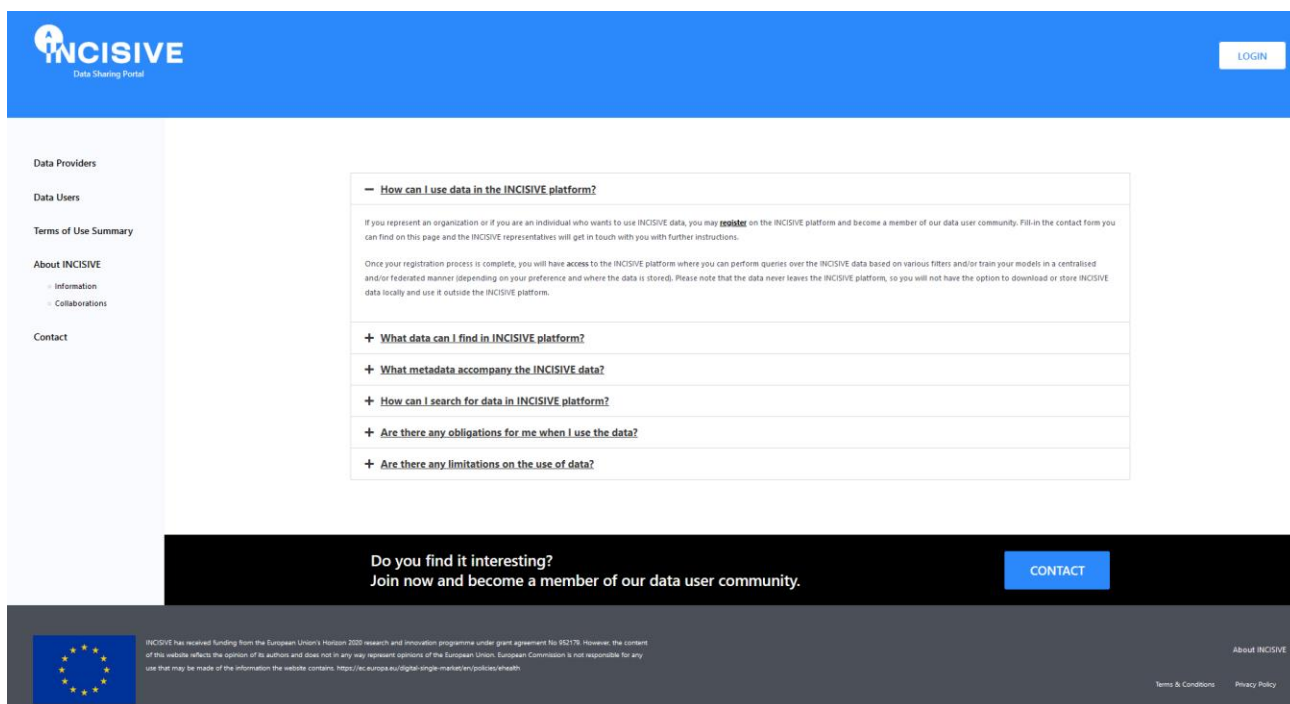
[CONTACT](#)

 INCISIVE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952179. However, the content of this website reflects the opinion of its authors and does not in any way represent opinions of the European Union. European Commission is not responsible for any use that may be made of the information the website contains. <https://ec.europa.eu/digital-single-market/en/policies/ehealth>

About INCISIVE

[Terms & Conditions](#) [Privacy Policy](#)

Figure 28. Openly accessible information for DPs.



INCISIVE
Data Sharing Portal

[LOGIN](#)

- Data Providers
- Data Users
- Terms of Use Summary
- About INCISIVE
 - Information
 - Collaborations
- Contact

How can I use data in the INCISIVE platform?

If you represent an organization or if you are an individual who wants to use INCISIVE data, you may **register** on the INCISIVE platform and become a member of our data user community. Fill-in the contact form you can find on this page and the INCISIVE representatives will get in touch with you with further instructions.

Once your registration process is complete, you will have access to the INCISIVE platform where you can perform queries over the INCISIVE data based on various filters and/or train your models in a centralized and/or federated manner (depending on your preference and where the data is stored). Please note that the data never leaves the INCISIVE platform, so you will not have the option to download or store INCISIVE data locally and use it outside the INCISIVE platform.

What data can I find in INCISIVE platform?

What metadata accompany the INCISIVE data?


How can I search for data in INCISIVE platform?

Are there any obligations for me when I use the data?

Are there any limitations on the use of data?

Do you find it interesting?
 Join now and become a member of our data user community.

[CONTACT](#)

 INCISIVE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952179. However, the content of this website reflects the opinion of its authors and does not in any way represent opinions of the European Union. European Commission is not responsible for any use that may be made of the information the website contains. <https://ec.europa.eu/digital-single-market/en/policies/ehealth>

About INCISIVE

[Terms & Conditions](#) [Privacy Policy](#)

Figure 29. Openly accessible information for DUs.

Moreover, two different forms for new DPs and DUs will enable the registration of new members of the INCISIVE data sharing mechanism.

Table 2. Registrations form for DP's and DU's.

Registration form fields	
Candidate Data Provider Form	Candidate Data User Form
Data Contributor Institution * Address * Primary contact person Name * Job position * Email * Phone Data Protection Officer Name * Email * Phone Other contact persons providing images if different from the Primary contact person Name Email Phone Are there any usage restrictions on this data? Description of data Modalities Body parts examined Number of patients	Details of the applicant Institution * Address * Website Primary Investigator Name * Job position * Email * Phone Google scholar page (or similar) Description of the research project Title of the project * Aims and research question * Explain why the data from INCISIVE platform is needed for your research project? * Do you have ethics approval? If yes, please provide details of your ethics approval. *

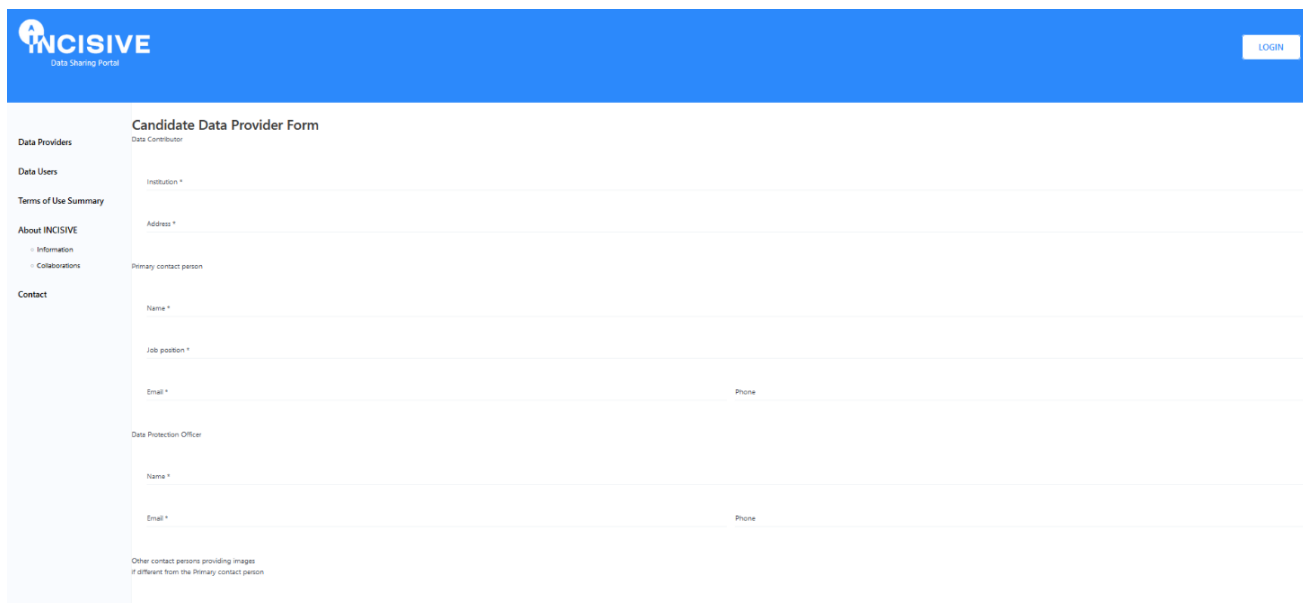


Figure 30. Registration form for DPs.

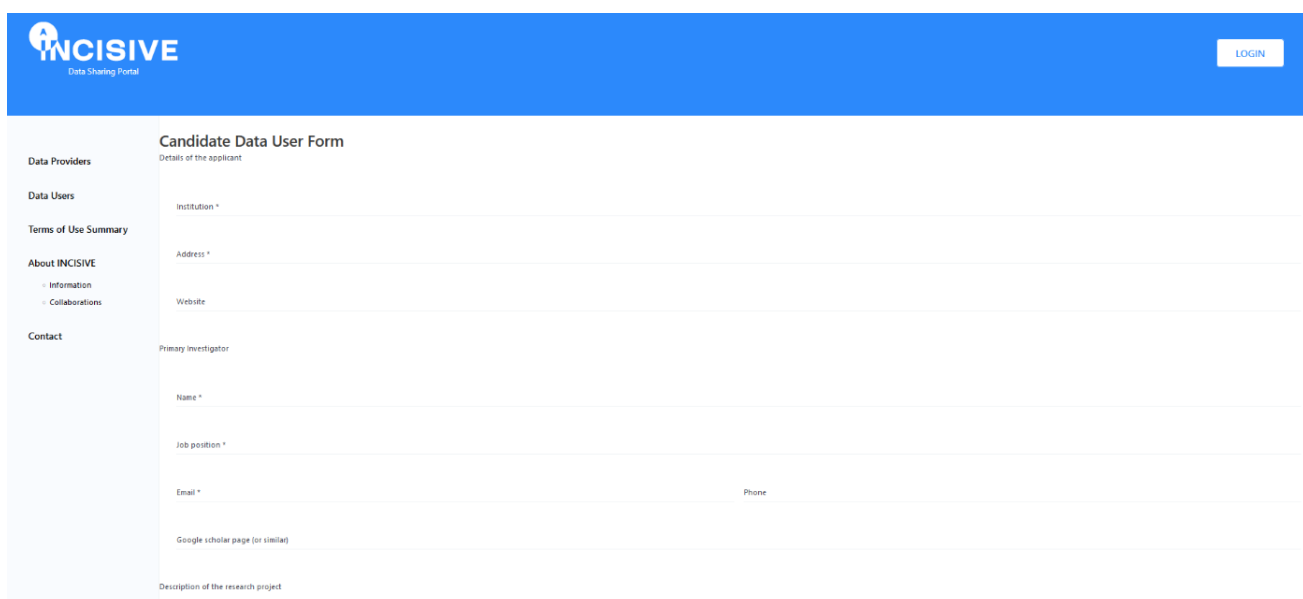


Figure 31. Registration form for DUs.

In order to enable the use of the services either for data sharing or use, a registered user shall first login into the system by clicking on the button on the upper right part of the screen. Then the user is re-directed to the INCISIVE login screen in order to get access to the repository.

A summary to the Terms and Conditions, information about the INCISIVE project along with its collaborations, a general contact form and links to the full Terms & Conditions and the Privacy Policy are also available.

6. Conclusions

This deliverable presents the third and final prototype of the INCISIVE federated repository of health images. It is describing all the functionalities regarding data management and storage, including all the latest updates since the previous iteration. The presented infrastructure and functionalities should be considered final, with all the required components in place and fully functional, reflecting the decisions that were taken by the consortium in an effort to satisfy all the requirements set since the beginning of the project, but also the ones that came up during the design and development stages.

The main updates documented throughout this document include the finalization of the data preparation process and tools used to achieve it, the expansion of the criteria for performing data searches both in central and federated nodes, the final version of the logging and auditing mechanism, in the form of a Transaction Tracker, as well as the utilization of blockchain to implement a reputation system for the AI algorithms, the final version of the data sharing portal with all the necessary functionalities for onboarding new members in the repository, and finally the latest updates on the UI for all the components.

With the creation of the INCISIVE pan-European repository of health images, the INCISIVE project aims to tackle the data availability issue, by creating a space for any organization or individual, in any European country, to not just share and safely store health data, but at the same time homogenizing the data and rendering it useful for a number of purposes, including research. All the components described in this document work together to achieve the aforementioned goal, whilst ensuring a high level of privacy and interoperability, providing the final users with all the necessary tools to either share their data, or use the existing data on the repository for research purposes, through an intuitive and user-friendly environment that guarantees a seamless experience.

References

- [1] D2.3: “INCISIVE Scenarios Definition” - INCISIVE - H2020
- [2] D2.5: “INCISIVE System Design - Final Version” – INCISIVE – H2020
- [3] D3.3: “INCISIVE Infrastructure - Final Version” - INCISIVE - H2020
- [4] D4.3: “INCISIVE AI Toolbox, Data Analytics and User Services - Final Version” - INCISIVE - H2020
- [5] D5.2: “INCISIVE Pan-European Repository of Health Images - Second Version” – INCISIVE – H2020
- [6] D6.2: “INCISIVE integrated Prototypes - Second Version” - INCISIVE - H2020
- [7] D8.5: “Preliminary Operational, Deployment and Sustainability Plan” - INCISIVE - H2020
- [8] FAIR Principles, <https://www.go-fair.org/fair-principles/>, last accessed: 28/07/2023
- [9] GDPR, <https://gdpr-info.eu/>, last accessed: 28/07/2023
- [10] PS3.15 - DICOM Standard - NEMA, 2023, <https://dicom.nema.org/medical/dicom/current/output/html/part15.html>, last accessed: 28/07/2023
- [11] Likert scale, <https://www.simplypsychology.org/likert-scale.html>, last accessed: 28/07/2023

ANNEX I – System Requirements Checklist

This ANNEX summarizes the final progress for the related to this deliverable system (and WP) requirements (functional (FSR) and non-functional (NF)). For each one the following status update options are available:

- **Fully Achieved:** whether the relates FS, NFS is achieved
- **Partially Achieved:** providing information on the progress as well as when/if this will be fully achieved
- **Not Achieved:** providing a justification why this was not achieved along with an estimation if this going to be partially of fully achieved within the project.

As already presented in the latest version of the design document the primary focus of the developments is to achieve the must have requirements. As seen below, most of the requirements that fall under the “Must” category, have been fully achieved, with the ones that were a work in progress in the previous iteration, now reached their full achievement.

Table 3. Non-functional System requirements.

Code	Title	Description	Metrics	Priority	Current Status	Justification
NF-FN-01	Participation	For participating in INCISIVE federated eco-system, a node must be set on premise	Available infrastructure on premises (4vCPUs and 16 GB of RAM)	Must	Fully Achieved	-
NF-DP-01	Metadata of datasets	Metadata should be well structured to support search options and advanced search option	Compliance to OMOP and HL7 FHIR standards	Should	Partially Achieved	HL7 FHIR is already supported, and while there have been significant improvements, some complex searches cannot be performed.
NF-AI-01	Performance	The jobs submitted should be executed within acceptable timeframe (robustness)	< 24 hours for a job to be executed	Must	Fully Achieved	Regarding the Federated search functionality.
NF-DS-01	Sharing of data	Data sharing must be performed based on legal and ethical norms following a well-	100% compliance legal and ethical requirements	Must	Fully Achieved	The data sharing mechanism has been developed and integrated

		defined semi-automated (digital) process				with the final prototype.
NF-OPS-01	SYS Modularity	System components must be built as modular and/or portable applications	Majority of components organized in modules, >90% of components built to be run in container environment	Must	Fully Achieved	Regarding the Federated/Central Storage components and functionalities
NF-OPS-15	SYS Interoperability	System edge node deployments must be interoperable	Nodes can be deployed in a variety of environments, at least Linux and Windows systems	Must	Fully Achieved	-

Table 4. Functional System requirements.

Code	Title	Description	Priority	Current Status	Justification
FSR-FN-01	FN Deployment	System must provide decentralized/distributed deployment within edge nodes of the system (Federated Nodes)	Must	Fully Achieved	-
FSR-FN-02	Update the data of a federated node	Data providers should be able to update their local data	Must	Fully Achieved	Data Providers can update their local data when data for an existing patient is ingested through the ETL tool.
FSR-FS-02	Search for medical data	Users should be able to discover medical data to train models with	Must	Fully Achieved	-
FSR-FS-03	Inquiry about the progress of a training taking place	Users should be able to find out what is the current status of the training process	Must	Fully Achieved	-
FSR-FS-03	Download a trained model	Users should be able to download a previously trained model	Must	Fully Achieved	-
FSR-FS-04	Terminate an ongoing training	User should be able to cancel the ongoing training of a model	Must	Partially Achieved	This functionality is not yet provided.

FSR-DP-01	High quality annotations	The system should provide the user an environment suitable for annotating DICOM images in a semi-automatic way	Must	Fully Achieved	-
FSR-DP-02	The data must be checked for quality issues	The system should be able to check the quality of the data uploaded and produce reports.	Must	Fully Achieved	-
FSR-DP-04	The data must be checked for quality issues	The user should be able to check the quality of the data and correct errors prior to data uploading	Must	Fully Achieved	-
FSR-ST-01	Storage of de-identified information	The system must be able to keep de-identified information	Must	Fully Achieved	-
FSR-ST-03	Storage of metadata	The system must be able to store meta data related with the de-identified information or the result of a process	Must	Fully Achieved	-
FSR-ST-04	Storage of annotated images	The system must be able to store files other than DICOM for images	Must	Fully Achieved	-
FSR-AI-02	Train a model	Users should be able to train a model using previously retrieved medical data	Must	Fully Achieved	-
FSR-AI-03	Asynchronous execution	The AI submitted jobs must be executed in an asynchronous way	Must	Fully Achieved	For the Federated Search functionality
FSR - VI-04	Visualization of results	The users should be able to access the results of the AI service in a visual form.	Must	Fully Achieved	
FSR - VI-05	Dataset search UI	There should be an intuitive query interface in the platform (simple/advanced search) for dataset search and discoverability	Must	Fully Achieved	-
FSR - SP-02	Data privacy	The system should allow the user to de-identify DICOM images	Must	Fully Achieved	-
FSR - SP-04	Auditing mechanism for tamper-proof logging	The system should provide a way of logging critical actions in an immutable way	Must	Fully Achieved	-

ANNEX II – Curation Script

It was deemed mandatory, that issues pertaining to the folder structure and the quality of the retrospective data, which had already been uploaded, needed to be addressed. To tackle these issues, a specialized curation script was developed and executed as a one-time measure. The primary objective of this script was to identify and, if possible, automatically rectify any issues that arose during the annotation and data uploading processes. In particular, the curation script was able to detect abnormalities in the folder structure of the uploaded data and extract a report for each data provider, in case manual correction was required. In most of the cases the curation script could automatically correct the issues related to wrong folder structure. The intervention of the data providers was necessary in situations where changes to the produced annotation files was evident.

The analysis of the retrospective data revealed that the majority of issues stemmed from a false folder structure, primarily caused by potential variances in adherence to the provided guidelines. In Figure 32 is presented the folder structure as it was proposed and agreed among the consortium. These discrepancies introduced inconsistencies and hindered the seamless integration of the data into the research framework. Consequently, the development of an automated correction mechanism became imperative to ensure the accuracy and reliability of the retrospective data.

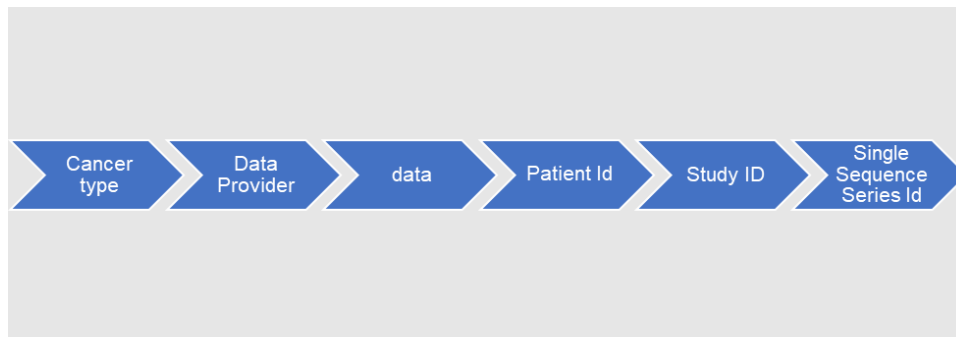


Figure 32. Proposed folder structure.

For cases where automatic correction was not feasible, a comprehensive report (Figure 33) was generated for each data provider. These reports detailed the specific manual corrections that needed to be undertaken by the respective data providers themselves. This approach ensured that healthcare professionals, possessing the requisite medical expertise and authority, were actively involved in rectifying issues related to the annotation process. Limitations in making corrections to issues directly associated with the annotation process, were acknowledged, thus necessitating the involvement of domain experts. By adopting this

meticulous approach, the integrity and dependability of the retrospective data were assured. This was of paramount importance, as it would enable AI developers to leverage the data for their research and development purposes with confidence.

```

Found incompatible number of DICOM files and NIFTI slices.
A medical expert should examine the specific annotation file and correct it.
Issue in file: /mnt/nfs/incisive3/disba/breast/007-020069/007-020069_FusPT_BL/Series-9951
Number of DICOM files: 255
Number of NIFTI files: 1
Number of NIFTI slices: 1019
Number of different DICOM sequences: 1

Found incompatible number of DICOM files and NIFTI slices.
The 4 dcm files should be separated into 4 different series folders. The annotation file should be placed together with the corresponding dcm file.
Issue in file: /mnt/nfs/incisive3/disba/breast/007-020070/007-020070_MG_BL/Series-71300000
Number of DICOM files: 4
Number of NIFTI files: 1
Number of NIFTI slices: 1
Number of different DICOM sequences: 4

Found incompatible number of DICOM files and NIFTI slices.
A medical expert should examine the specific annotation file and correct it.
Issue in file: /mnt/nfs/incisive3/disba/breast/007-010012/007-010012_US_BL/Series-1
Number of DICOM files: 14
Number of NIFTI files: 1
Number of NIFTI slices: 1
Number of different DICOM sequences: 1
  
```

Figure 33. Typical example of a generated report containing the precise issues and the proposed solution.

The main issue that demanded rectification was associated with the folder structure, specifically concerning the guideline that stipulated each Series folder should encompass a solitary sequence of DICOM files accompanied by its corresponding annotation NIFTI file, if applicable. It was crucial to identify cases where multiple sequence Series folders coexisted for a specific cancer type and image modality, and subsequently segregate them into distinct, single sequence Series folders.

To address this particular issue, an initial comprehensive exploration of the database was undertaken. The analysis revealed that the aforementioned problem was predominantly observed in Series folders linked to breast MRI, breast MG, and colorectal MRI. These specific modalities exhibited a higher incidence of multiple sequence Series folders, which necessitated the implementation of corrective measures to ensure adherence to the prescribed folder structure guidelines.

The implementation of the curation script yielded positive outcomes by successfully separating non-annotated multi-sequence Series folders into their respective single sequence counterparts. Additionally, the script demonstrated a high success rate in correctly dividing annotated multi-sequence Series folders. However, a notable challenge emerged in cases where annotation files were exclusively available for specific DICOM sequences within a multi-sequence Series folder. The lack of explicit information indicating the correspondence between DICOM sequences and their respective annotation files made it unfeasible to automatically rectify this particular issue.

As a result, a comprehensive report was generated to document the occurrences of this issue. The responsibility for resolving this challenge was then assigned to healthcare professionals who possessed the necessary expertise in the medical domain. These professionals were tasked with manually examining the affected multi-sequence Series folders and determining the appropriate alignment between DICOM sequences and their corresponding annotation files. This approach acknowledged the limitations faced by technical partners in making direct corrections to annotation-related issues that required medical expertise. By involving healthcare professionals, the process ensured that the correct alignment between DICOM sequences and annotations was established, thus upholding the accuracy and reliability of the curated retrospective data.

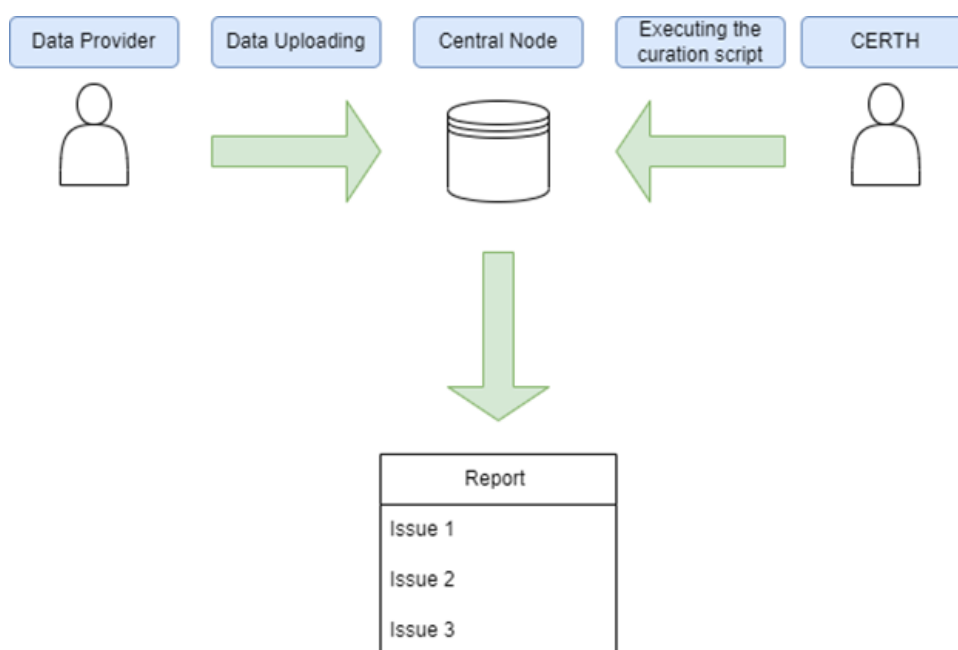


Figure 34. Report generation after a data provider uploads new data to the central node.

Furthermore, the curation script successfully identified cases where annotation files were found outside of their corresponding Series folders. In most instances, when a clear correspondence existed between the Series folder and the NIFTI annotation file, the script was able to automatically resolve this issue. However, complications arose when the NIFTI file was discovered in a directory containing multiple Series folders, making it impossible to definitively determine the specific Series folder to which it belonged.

To address this challenge, the expertise of a medical professional was indispensable. These medical experts were responsible for meticulously examining the medical files and precisely assigning each NIFTI file to its appropriate Series folder. Through their domain knowledge and

expertise, they were able to accurately associate the annotation files with the respective DICOM Series folders, ensuring the integrity and consistency of the curated data. In addition to the aforementioned issues, the curation script also identified and addressed empty Series folders within the dataset. These empty folders were deemed redundant and were consequently removed from the dataset, ensuring the overall cleanliness and organization of the curated data.

Moreover, the script examined Series folders where inconsistencies were observed between the total number of DICOM files and the corresponding NIFTI slices. These discrepancies were reported for further investigation, aiming to determine whether errors had occurred during the annotation procedure. It was imperative to verify the accuracy and reliability of these Series folders, as any discrepancies could potentially impact subsequent analyses and research outcomes.

During the curation process, it was discovered that a significant number of DICOM and NIFTI files lacked essential header information or exhibited invalid dimensions. As a result, these files could not be effectively examined and incorporated into the curated dataset. However, to ensure a comprehensive and thorough approach, all files that failed to be successfully read were diligently documented. This documentation serves as a reference for future inspection and potential re-upload, enabling the identification and resolution of issues associated with these files.

By addressing empty Series folders, reporting discrepancies, and documenting problematic files, the curation process aimed to enhance the overall quality and reliability of the retrospective data. This systematic approach ensures that the curated dataset is robust, accurate, and suitable for utilization in AI development and medical research endeavours.

The manual correction process, undertaken by the data providers, was being closely monitored to ensure the reliability of the results, while relevant information was provided related to the extracted reports and the appropriate resolution of identified issues. This monitoring process guarantees effective collaboration, facilitating the timely and accurate correction of any remaining issues.

Looking forward, the curation script was also systematically executed for prospective data uploading. This systematic approach enables the identification and resolution of any issues related to the prospective dataset, ensuring the ongoing accuracy and dependability of the data. By applying the curation script to prospective data, the project can proactively address any potential issues and maintain a high standard of data quality throughout the research project.

Once the manual correction process was finalized, the automatically corrected data was merged with the data that has been rectified and re-uploaded by the healthcare

professionals. This merging process (Figure 35) aimed to consolidate the corrected data and ensure its seamless integration.

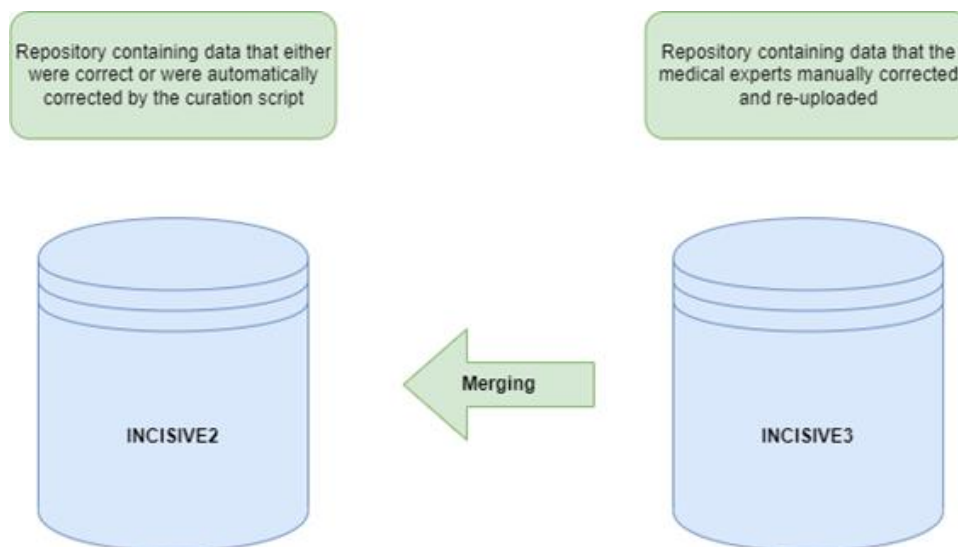


Figure 35. Merging data corrected by the data providers with data corrected by the curation script.

By merging the automatically corrected data with the manually rectified and re-uploaded data, the creation of a unified dataset that incorporates all the necessary corrections and adjustments was achieved. This consolidated dataset embodies the collective efforts of both automated and manual interventions, resulting in a comprehensive and refined dataset. The merging process involves careful data integration techniques to reconcile any potential conflicts or inconsistencies between the automatically corrected data and the manually corrected data. Appropriate strategies were applied to ensure the accuracy, consistency, and reliability of the merged dataset.

Ultimately, the merged dataset serves as the foundation for subsequent analysis, AI development, and research activities, enabling researchers and AI developers to leverage a high-quality and dependable dataset for their work.