Intelligent Ecosystem to improve the governance, the sharing, and the re-use of health Data for Rare Cancers

D1.2 Data Management Plan

Distribution List

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

Revision History

Revision no.	Date of Issue	Author(s)	Brief Description of Change
0	02/12/2022	E. Martinelli (INT)	ToC
0.1	02/12/2022	E. Martinelli (INT)	ToC
0.2	19/12/2022	E. Gaeta (UPM	Added contributions
0.3	30/12/2022	A. Trama (INT), E. Gaeta (UPM)	Added contributions
0.4	12/01/2023	P. Zampognaro (ENG), A. Almeida (UDEU)	Added contributions
0.5	24/01/2023	CERTH, UU, IKNL	Added contributions

0.6	19/12/2022	E. Gaeta (UPM)	Added contributions
0.7	31/01/2023	E. Martinelli (INT)	Added contributions
0.8	01/02/2023	G.Cangioli	Added contributions
0.9	05/02/2023	A. Trama (INT)	Added contributions
1.0	08/02/2023	I. Drympeta (CERTH)	Added contributions
1.1	09/02/2023	S. Ziegler (ECCP)	Added contributions
1.2	10/02/2023	E. Martinelli (INT)	Integrated version
1.3	14/02/2023	E. Martinelli (INT	Integrated revisions by APHP
1.3	28/02/2023	V. Tsiompanidou (ECCP)	Peer-review
1.4	01/03/2023	E. Martinelli (INT)	Final versions including peer.re'iewer's recommendations

Addressees of this document

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

TABLE OF CONTENTS

1 Data Summary 7
1.1 Purpose of use and re-use of data in IDEA4RC 7
1.2 Data types and formats 7
2 FAIR Data 9
2.1 Making data findable, including provisions for metadata 10
2.2 Making data accessible 11
2.2.1 Data Repository 11
2.2.2 Data 12
2.2.3 Publications and Intellectual Property Rights Policy 14
2.2.4 Data accessibility and data access procedures 15
2.2.5 Metadata 15
2.3 Making data interoperable 16
2.4 Increase the data re-use 18
3 Other research outputs 20
4 Allocation of resources 23
5 Data security 25
6 Ethics 27
6.1 Regulatory framework 27
6.1.1 Applicable regulations 27
6.1.2 Upcoming Regulations 28
6.1.3 Ethics, personal data protection and management 29
6.1.4 Data Protection Coordination Board 29
6.1.5 Clarification on the Responsibilities 29
6.1.6 Ethical Requirements 30
6.1.7 IDEA4RC Policy 30
6.1.8 Data Sharing Policy 31
7 Other issues 32

LIST OF TABLES

Table 1. Example of costs for different types of research products 23

LIST OF FIGURES

Fig. 1. Schematic presentation of the TEHDAS User Journey 9
Fig. 2. Scheme of the IDEA4RC trusted distributed repositories 11
Fig. 3. The IDEA4RC FHIR Implementation Guide 16
Fig. 4. Interoperability standards for data and AI models in IDEA4RC 17
Fig. 5. Data and other research outputs management and sharing 20
Fig. 6. EOSC Onboarding Process (adapted from https://eosc-portal.eu/) 21
Fig. 7. EOSC information model (adapted from https://eosc-portal.eu/) 22
Fig. 8. Example of workflow for the publishing process of different research products 22
Fig. 9. IDEA4RC relies on 5G Concepts 25

Abbreviations and definitions

Acronyms

CA	Consortium Agreement
DoA	Description of Action – Annex I to the Grant Agreement
EC	European Commission
EHDS	European Health Data Space
EU	European Union
FHIR	Fast Healthcare Interoperability Resource -is an interoperability standard developed by HL7 (the Health Level 7 standards organization) designed to enable the exchange of healthcare data electronically between different systems in the healthcare industry.
GA	General Assembly
KPI	Key Performance Indicator
QA	Quality Assurance
QP	Quality Plan
QC	Quality Control
RAM	Risk Assessment Matrix
RC	Rare Cancers
SC	Steering Committee
SSH	Social Sciences and Humanities
WP	Work Package

1 DATA SUMMARY

1.1 Purpose of use and re-use of data in IDEA4RC

The main objective of the IDEA4RC project is to establish a Data Space for rare cancers (RC) that will make possible the re-use of existing multisource health data across European healthcare systems.

Data will be re-used to advance research, increase knowledge, ameliorate quality of care and access to optimal treatments for patients with RC.

The main problem of RCs (those with an incidence < 6/100,000) is by definition the low number of cases, which makes it difficult to carry out clinical trials and build clinical evidence and therefore makes clinical management more complex. This limiting factor can be overcome only through large collaborations exploiting data already available within networks specializing in rare cancers, that pool knowledge and data together. For this reason IDEA4RC is developing and establishing an ecosystem of tools that enable federated data use and are able to manage data access and reuse conditions and permissions through a privacy-by-design framework embedded in the ecosystem.

We will re-use the data, generated during the clinical routine, in 11 expert centers of the European reference network on rare adult solid cancers (EURACAN https://euracan.eu/). We do not envision an ad hoc data collection during the project.

We will re-use data already available in electronic format (e.g., electronic health records, clinical registries, administrative datasets etc.) both structured and unstructured (e.g., from pathology, radiology, visits, discharge etc. referrals). Unstructured data will be processed to extract structured information to facilitate data analysis and reuse. Biological data will not be collected nor used.

The IDEA4RC data space will integrate different types of data from RC expert centres. It will be a unique opportunity to access a large, high quality, standardized and highly powerful data platform on RC. It will be useful to:

1) researchers and clinicians of all the EURACAN centers and/or interested in RC (the rare cancers community is much wider than the 11 centres involved in the project);
2) national or European competent authorities (e.g., European Medicines Agency);
3) patients and caregivers;
4) data scientist and statisticians; 5) scientific professional societies;
5) pharmaceutical companies.

1.2 Data types and formats

In IDEA4RC the main data that will be collected and/or processed are related to clinical data from electronic health records (EHRs) and from unstructured (i.e., text) referrals such as from pathology, diagnostic imaging, discharge letters and medical prescriptions, from and case report forms (eCRF) during clinical trials and data from rare cancer registries. in the course of the project, no genomic data will be collected, nor diagnostic images (DICOM format). As the project progresses, additional information will be provided on the precise datasets involved in IDEA4RC, as will be reported in future iterations of the Data Management Plan.

In order to improve the interoperability, use and re-use of data managed during the project by the pilot sites, the standard HL7 FHIR will be adopted and a novel FHIR implementation guide for head and neck cancer and sarcoma will be produced.

Anyway, existing initiatives that are already working in the specification of a common data model for cancer such as minimal Common Oncology Data Elements (mCODE[1]) or Interoperability and data sharing of clinical and biological data in Oncology (OSIRIS[2]) will be deeply analyzed, adopted and extended if they comply with the requirements found for IDEA4RC. Other related initiatives not based on FHIR such as the adoption of the Observational Medical Outcomes Partnership -- Common Data Model (OMOP-CDM[3]) for Rare Cancers will be further explored.

From a preliminary study we have made with the participating centers, we have found that our data are mainly text data, no -omic data and images will be available in this early phase of the project. This fact reduces the size of the data that will be treated and the storage resources that need to be allocated.

On the other hand, as we have a big amount of data in text format that need to be analyzed, we expect that High Performance Computer (HPC) environments equipped with Graphics Processing Unit (GPU) need to be used.

A more precise estimate of the size of the datasets which will be generated and managed by the project will be available in a later stage, when the data infrastructure will be developed and populated. This information will be reported in the updated versions of the present document.

[1] mCode, FHIR implementation guide, http://hl7.org/fhir/us/mcode/ , last access Dec. 2022
[2] OSIRIS, FHIR implementation guide, https://fhir.arkhn.com/osiris/ , last access Dec. 2022
[3] OMOP-CDM, https://www.ohdsi.org/data-standardization/, last access Dec. 2022

2 FAIR DATA

IDEA4RC will deliver a digital platform to enhance the use and reuse of clinical data in trusted research environments (IDEA4RC FHIR capsules) that will enable privacy-enhanced preserving data analysis for research study on head and neck cancer and sarcoma. The main users of the IDEA4RC platform will be:

1. the researchers that would analyze data coming from different centers that are collecting data on these rare diseases.
2. On the other hand, there will be the medical centers that would aim to avoid to researchers to inspect patient-level data that include sensible and personal information of their patients. But, at the same time they want to enable researchers to execute the research studies to increase our knowledge on rare cancers.

The aims of IDEA4RC show commonality with the European Health Data Space (EHDS) regulation and with the ongoing works of the associated project "Towards European Health Data Spaces" (TEHDAS. For instance, just to mention a few relations: the IDEA4RC researcher is what EHDS mentions as Data User, the medical centers that process and control the data are defined Data Holders in EHDS regulation, the IDEA4RC federated platform is what the EHDS identifies as Data Access Body, the IDEA4RC FHIR Capsules are what the EHDS recognizes as Secure Processing Environments.

If we look at the TEDHAS user journey (Figure 1), we can establish clear relationships among the phases described in the journey and the components within the IDEAR4RC platform that are designed to implement the services associated with each phase of the TEHDAS user journey.

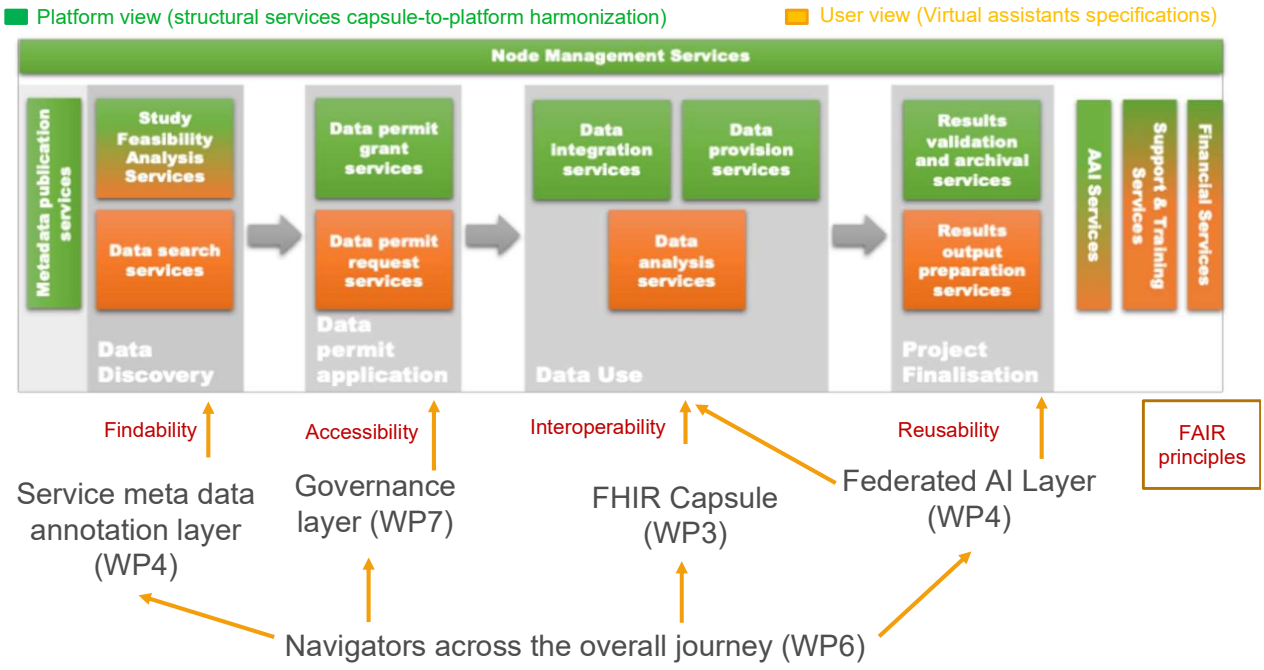

Fig. 1. Schematic presentation of the TEHDAS User Journey

In Figure 1 the relations among TEHDAS phases including services of the user journey, Findability Accessibility, Interoperability, Reusability (FAIR) principles and IDEA4RC Work Packages (WP) and components are highlighted:

- Data Discovery phase associated with the Findability principle is addressed in IDEA4RC by the Service meta -annotation layer provided by WP4;

- Data permit application phase associated with the Accessibility principle is provided in IDEA4RC by the Governance Layer implemented in WP7;
- Data use phase associated with the Interoperability principle is aligned with the concept of FHIR capsule of IDEA4RC, devoted to the implementation of interoperable data sharing and privacy preserving processing environment (WP3) including the advanced services for data cleaning, data augmentation and federated Artificial Intelligence (AI) implemented in the task 4.3 of WP4;
- Project finalization phase associated with the Reusability principle that is part of the federated IDEA4RC federated platform developed into WP4.

Furthermore, across all the user journey described in Figure 1, IDEA4RC project will develop virtual assistants with the aim to support users in enhancing their experience along all the phases of the journey.

2.1 Making data findable, including provisions for metadata

Making data findable is one of the most important aspects of the metadata annotation layer.

To ensure that data is findable, the IDEA4RC project will ensure that the data model definition in T5.1 and the metadata definition from T2.4 are harmonized. In this manner, the variables and concepts that need to be found are clearly defined. These variables are specified by experts and stakeholders of the IDEA4RC project in task 8.1 of WP8 and task 2.1 of WP2.

Finally, the service metadata annotation layer will ensure that the concerned variables are constantly marked to be findable and searchable in a constant process, summarizing the status in terms of quality of each of those variables. Last, data will also be annotated following the FAIR[4] principles where needed so it can be found.

As previously mentioned, metadata will follow the FAIR principles. In this regard, the findability and accessibility of data will be guaranteed in cases where legal aspects do not compel privacy. Data subject to GDPR will be managed according to National and Institutional regulations in compliance with GDPR. will apply the Data Catalog Vocabulary (DCAT[5]) to the IDEA4RC data, using a standard model and vocabulary that facilitates the consumption and aggregation of metadata from multiple centres.

Metadata of deposited data must be open under a Creative Common Public Domain Dedication (CC 0) or equivalent (to the extent legitimate interests or constraints are safeguarded), in line with the FAIR principles (in particular machine-actionable) and provide information at least about the following:

- datasets (description, date of deposit, author(s), venue and embargo);
- Horizon Europe or Euratom funding;
- grant project name, acronym and number;
- licensing terms;
- persistent identifiers for the dataset, the authors involved in the action, and, if possible, for their organisations and the grant.

Where applicable, the metadata must include persistent identifiers for related publications and other research outputs.

[4] FAIR principles, https://www.go-fair.org/fair-principles/, last access Jan. 2023
[5] Data Catalog Vocabulary (DCAT), https://www.w3.org/TR/vocab-dcat-2/, last access Jan. 2023

The search keywords, terms and process will be further designed and accorded in the co-creation process that will take place in T.2.1. Nevertheless, the project further anticipates that the searchable metadata vocabulary/standards will be publicly available.

Metadata will be indexed and accessible through the Navigator tool provided by IDEA4RC ecosystem. As previously discussed, the IDEA4RC project will follow the FAIR principles and use DCAT vocabulary to support machine readability, data discovery and findability.

2.2 Making data accessible

Accessibility of the data will be realized via the governance layer of IDEA4RC. The governance layer, which is to be implemented in WP7, will enable the data permit requests, as well as the negotiation between bilateral data agreements and the patient consent and data altruism consent to (re-)use of data. Data sovereignty will also include usage control of the data beyond the usual access control, leveraging usage policy management mechanisms compliant with the International Data Spaces[6]. The tools implemented will use open, free, and standardized communications protocols, such as HTTPS and MQTT, for the access to the data ensuring security and privacy.

2.2.1 Data Repository

Within IDEA4RC the data will not be deposited into repositories but they will be made available within "trusted research environments" called FHIR capsules that are systems aligned with the concept of "secure processing environments" that EHDS regulation proposal mentions in the article 50 [7].

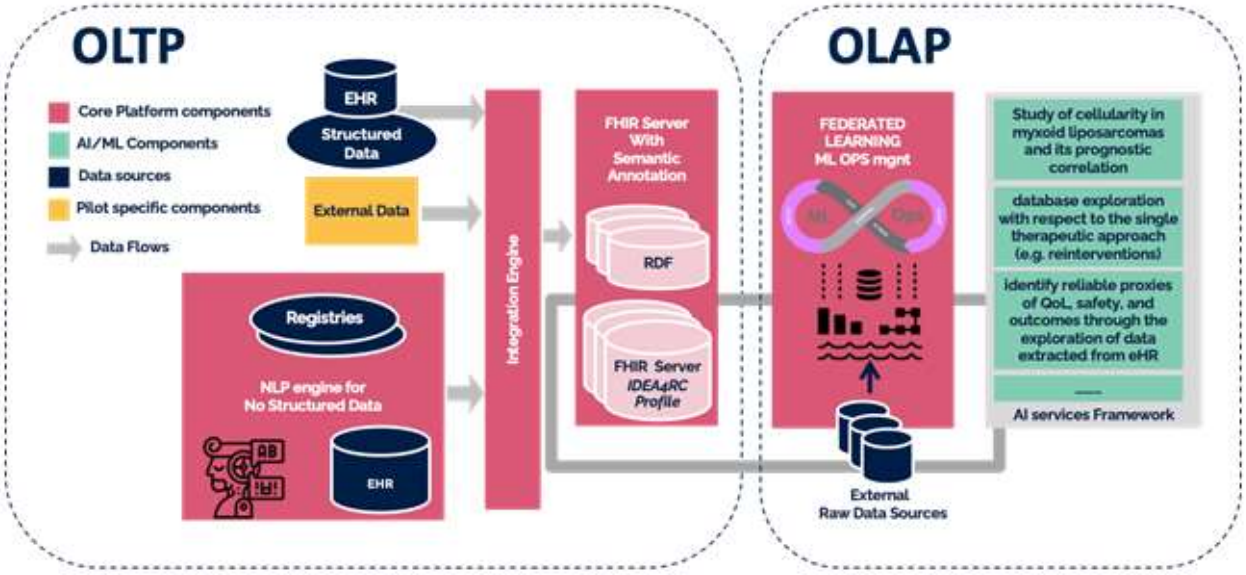

Fig. 2. Scheme of the IDEA4RC trusted distributed repositories

The FHIR capsule will provide 2 main functionalities:

[6] Steinbuss S. et al. (2021): Usage Control in the International Data Spaces. International Data Spaces Association. https://doi.org/10.5281/zenodo.5675884

[7] EHDS regulation proposal, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0197&from=EN , last access Dec. 2022

- on line transactional processing system (OLTP), each capsule will be able to store, collect and inject data from different sources into a common interoperable format based on HL7 FHIR, it is expected the publication of a IDEA4RC FHIR implementation guide as human readable set of resources of the common data model created along the project;
- on line analytics processing system (OLAP), each capsule will be able to locally process and analyze the data when the users are requesting to access to personal and sensitive data stored into the capsule. The data in the capsule will be pseudonymized and will not be disclosed outside the data holder's premises, but will be analyzed in federated way. The result of data processing and analysis will be provided in anonymized format.

Depending on the needs of each data holder, the data included into each FHIR capsule could have persistent identifiers. Such needs will be expressed as requirements during the execution of tasks 2.4 and 5.1 where requirements on data and meta data will be extracted.

Clinical partners of the IDEA4RC Consortium are aware of the proposed data infrastructure and of the need to make available internal data storage resources as well as of the need to implement the necessary IDEA4RC components that allow federated data access under the data governance provisions and procedures agreed and defined by the data holders in the frame of IDEA4RC co-creation work package (WP2).

2.2.2 Data

The project ambition is to make data openly available, however, barriers that exist among data holders, starting from IDEA4RC partners, for sharing, exploring, and re-using their data have to be explored to ensure data openness. To this extent, a co-creation process (Task 2.1 of WP2), will be enacted with the different stakeholders involved in RC data provision, sharing and reuse, to understand governance, processes and rules for data sharing as well as to elicit incentivization processes that can be enacted within the RC data ecosystem to make data openly available. Additionally, a Data Protection Coordination Board (DPCB) including each partner's Data Protection Officer (DPO) will ensure GDPR compliance by monitoring what data is processed and how; how data is transferred, safeguarded, etc. within the project always with the intention of properly addressing legal obligations to make the data accessible as openly as possible. A Data Protection Impact Assessment will also be discussed by the DPCB leveraging on risks identified by the data holders themselves. Finally different contractual mechanisms will be discussed based on GATEKEEPER experience; split in 3:

- General set of obligations for *all* partners depending on their role; closely aligned with the GDPR and SCCs (general framework)
- Complementary agreements
- 3rd party obligations (for non-partners but needs to respects certain obligations) → Terms of Use

Co-creation approaches are enacted to involve data providers (Data Holders) in the definition of the limitations and procedures required to make data re-usable for scientific research and clinical purposes. User stories describing use scenarios identified and described by pilot users and IDEA4RC stakeholders, will be the first step of this process.

From the user stories surveys input will be elicited that is used in the organisation of co-creation activities, drawing on a focus on values and looking for example at potential tensions, which need to be further explored, such as how do values like data privacy differ for different partners and stakeholders

and what potential challenges these raise for data sharing[8]. In the co-creation work we aim to further enlarge and / or finetune the IDEA4RC ecosystem map, which can be considered a dynamic context. However, the aim is towards an as comprehensive as realistically possible idea of potential users and / or stakeholders. With relevant stakeholders involved and drawing on input from the collected user stories, further reflection can be organised on underlying values and valuation practices (cf. valuation approach). For example, by co-creating future data use, sharing and protection scenarios. Connections will be made to previous work and related work in other projects.

In short, the co-creation work enables to reflect on potential valuation practices, and to co-create future scenarios. As the field is dynamic and both context and actors involve may change, within the project an opportunity for ongoing reflection and co-creation beyond project duration would ideally be developed.

IDEA4RC intends to adopt and implement a dual strategy with regards to data, that will be focusing on the following main goals:

1. Protecting personal data and complying with the applicable regulations, including with the General Data Protection Regulation (GDPR[9]) and the Medical Devices Regulation (MDR[10]), while also considering upcoming regulations such as the Regulation on the European Health Data Space (EHDS[11]).

Within this context, IDEA4RC is bound to follow the data protection principles enshrined in the GDPR, namely the principle of personal data protection by design and by default, as well as the data minimisation principle. In view of this, where possible, personal data shall not be shared with other partners or third partners, but they shall be anonymised implementing all possible technical and organisational measures to prevent de-anonymisation.

Given the sensitivity of the data that may be relevant for the implementation of the project, additional safeguards will be put in place to ensure data subjects remain protected and their privacy is respected at all times.

Of course, in compliance with the GDPR, each partner may act as a data controller and/or data processor for personal data processed in the context of the project and, thus, each partner has a legal responsibility to comply with the applicable regulations.

However, as the data processing of the partners may also result in ethical issues and reputational risks for the project, even compromising it as a whole, it has been agreed to establish a Data Protection Coordination Board (hereafter 'DPCB') that will be in charge of coordinating and aligning the data processing of all partners with common rules and policies. Among the DPCB 's responsibilities lies the drafting of a concrete Data Sharing Policy that will regulate in detail the data sharing activities of the partners, not only among the Consortium but also beyond it, the procedures to be followed, as well as any safeguards that need to be implemented. More information on the Data Sharing Policy is provided in Section 6 of the present deliverable.

[8] see for example GATEKEEPER Trust Framework; 2021.

[9] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) 2016 (OJ L).

[10] European Parliament and Council of the European Union, REGULATION (EU) 2017/746 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU 2017.

[11] European Commission, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space 2022.

2. Maximising the availability of data for secondary use for research, in accordance as well with the Open Science principle. Open Science lies at the heart of the European Union's agenda, as it paves the way for innovation and societal evolution, promoting open access to data, publications, software etc.

IDEA4RC is focusing on ensuring compliance with the Open Science principle, in line, as well, with the UNESCO Declaration. In particular, as per 2021 UNESCO's Recommendation on Open Science[12], IDEA4RC aims at having its data meet the following conditions where possible:

a. Available in a timely manner,
b. Through a user-friendly format,
c. Human and machine-readable,
d. Actionable,
e. In accordance with the principles of good data governance, stewardship, the FAIR principles,
f. Supported by regular curation and maintenance[5].

Similarly, any licensing chosen for the datasets developed in the course of the project will, at the maximum degree possible, be providing other parties the right to user, access, modify, expand, study, create derivative works and share the software and its source code, design or blueprint. As far as possible, IDEA4RC will be sharing their datasets in open access repositories, such as the European Open Science Cloud (hereafter 'EOSC'), as will be further described below.

2.2.3 Publications and Intellectual Property Rights Policy

The Consortium Agreement lays down the foundation for the Intellectual Property Rights (hereafter 'IPR') Policy, providing the general framework regarding IPR ownership and exclusive rights of the results produced.

Based on the above, it is explicitly stated that IPR belongs, by default, to the partner who developed the innovation or content generated in the context of the project. If more partners have participated in the development of the same IPR, they are required to proceed to a written agreement describing their rights and obligations deriving from said IPR, as well as the terms and conditions of their exploitation.

In particular regarding publications, and in order to respect and uphold the IPR of the partners, the Consortium has agreed that all partners are due to give prior notice notice of any planned publication to the other Parties at least 45 calendar days before the publication as established in the Consortium Agreement Art. 8.4.2.1 so that any partner can raise any objections. The main author shall also share with the Consortium their draft publications before their publication, and at least 7 days prior, in order to allow Consortium partners to request reformulations.

The above-described IPR Policy will be complemented by Task T11.4 regarding the Intellectual Property Rights Agreements that will be reported in D11.4 Launch Plan and IPR Agreement plan and templates. In the context of said task, the IPR developed will be mapped, identifying the respective owners in each case and suggesting the most suitable IPR protection strategy.

In addition to the above and in line with the Horizon Europe policy and guidelines, all partners are due to maximise the exploitation of their research and innovation results, not only within the context of the project but also beyond it. In order to abide by this commitment, an IPR agreement will be drafted in the

[12] https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

context of the abovementioned T11.4, with the aim of facilitating the transition of IDEA4RC to a self-sustained operation after the end of the project.

2.2.4 Data accessibility and data access procedures

In IDE4RC data will be accessible through the FHIR standard.

Data will be shared on the basis of a formal Data Sharing Policy that will be drafted and maintained by the Data Protection Coordination Board, as mentioned above and as will be analysed below more in detail. Data that is compliant with said Data Sharing Policy may be shared with the partners, and, where applicable, with third parties not only for the duration of the project but also after its end.

Based on the above, each partner is responsible for maintaining their datasets available for at least 10 years after the end of the project, provided they are exempt from any personal data. As briefly explained, according to the GDPR, anonymised data do not constitute personal data and, thus, are not subject to personal data protection provisions. Of course, any such data may be adequately protected so as to prevent their de-anonymisation.

In addition to the above, partners will be encouraged to leverage the EOSC in order to make the datasets that will be produced available to the scientific community. The EOSC provides an optimal environment for researchers, innovators, companies and citizens to be given the opportunity to publish, find and reuse data, tools and services for research, innovation and educational purposes. As such, IDEA4RC may benefit from the solutions provided by the EOSC aiming at a high-level dissemination of the work performed, granting access not only to partners but also to the scientific community.

Established, open standards, such as OpenID Connect, will be used for user authentication where needed. Verifying user identity will follow the best practices as proposed by the Open Web Application Security Project (OWASP)[13], such as using generic error messages and storing passwords securely.

When a user wants to access the FHIR capsule, owned, deployed and managed by data holders and federated into the IDEA4RC platform, he/she will have to handshake with the Governance Layer of the IDEA4RC platform a set of access policies. The Governance layer will decide if a data user has the necessary credentials to access data into a FHIR capsule or to submit a data processing request on one or more FHIR capsules.

The access policies the Governance layer will be able to provide to data users, will be agreed upon among the Governance Layer and medical centres (FHIR Capsule holders) during the ecosystem creation phase and a digital contract will be implemented into the Governance layer to facilitate the process and the agreements needed for the data access.

2.2.5 Metadata

Metadata should be in line with the FAIR (Findable, Accessible, Interoperable, Reusable) principles, in particular, it should be machine-actionable (machine readable, and automatic computer processing can extract information from the metadata attributes ensuring a cross-linking between different research outputs) and follow a standardised format, in line with community standards, and should provide rich information on the publication/data (author(s), publication title, date of publication, publication venue); Horizon Europe or Euratom funding; grant project name, acronym and number; licensing terms.

[13] https://cheatsheetseries.owasp.org/cheatsheets/Authentication_Cheat_Sheet.html accessed 06-02-2023.

All the metadata will be open and CC0 licensed except in the cases, if they exist, where the publishing of metadata could lead to data deanonymization.[14]

As the main objective of IDEA4RC is to create and sustain a data ecosystem for rare cancers, based on data held at EURACAN Centers of Excellence, data and metadata will remain available and findable indefinitely. The specific conditions under which data and metadata can be accessed will be governed by Centers (ensuring data sovereignty and compliance with legal and ethical requirements) through the IDEA4RC data governance layer, to be developed in WP7. Such conditions may change along time, according to Centers' decisions.

Any relevant software needed for the implementation of the system will be provided as open source are freely available to provide the necessary environment to anyone that want to join the IDEA4RC open source community, anyone that want to adopt the IDEA4RC approach outside the project, anyone that want to review and test the software for deep validation and/or certification of the systems.

2.3 Making data interoperable

IDEA4RC is implementing interoperability agreeing a common implementation independent data model (also known as logical model, data set), defining, where appropriate, the reference terminologies[15] to be used for the coded elements. All this will be used, into each HL7 FHIR capsule, as basis to map the registries and EHRs data to the agreed HL7 FHIR profiles and reference value sets[16]. In this respect, an IDEA4RC FHIR implementation guide[17] is expected to be publicly published by the project collecting the formal representation of the agreed data models as HL7 FHIR logical models; the HL7 FHIR profiles to be used, including the reference vocabularies; and how the logical model map to profiles.

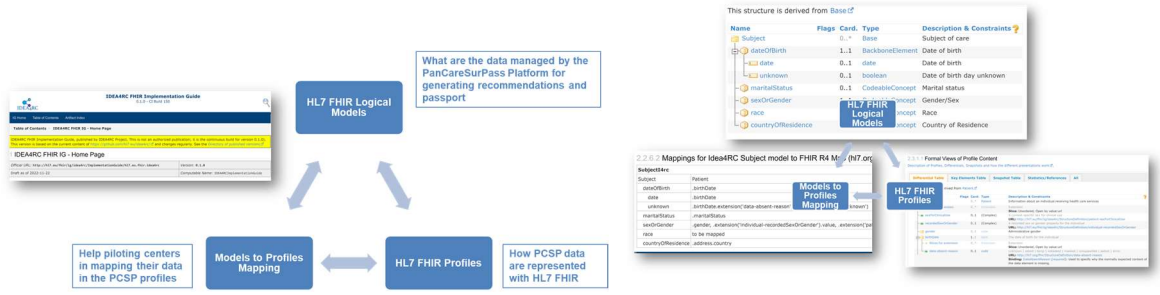

Fig. 3. The IDEA4RC FHIR Implementation Guide

HL7 FHIR interoperability relies on several aspects. From data representation point of view, it standardizes a first level basic ontology collecting a set of standard resources. Each HL7 FHIR "Resource" represents a health concept, for example a Condition, a Patient, a medication administration. Furthermore, it allows to support human and computable interoperability, allowing to bind the coded

[14] Paul M. Schwartz & Daniel J. Solove, The PII Problem: Privacy and a New Concept of Personally Identifiable Information, 86 N.Y.U. L. REV . 1814, 1843 (2011); Scott Berinato, There's No Such Thing as Anonymous Data, H ARV. BUS. REV . (Feb. 9, 2015), https://hbr.org/2015/02/theres-no-such-thing-as-anonymous-data [perma.cc/YF4C-6S6C] ("Anonymization . . .is 'inadequate' and ultimately doomed to fail with large metadata—the kind of publicly available big data that so many companies are tapping into.").

[15] As appropriate, the reference value set will be published with the HL7 FHIR Implementation Guides to allow their reuse, refinement or extension.

[16] As needed, formal vocabulary maps will be specified to map possible locally used value sets and those agreed for the data sharing.

[17] A HL7 FHIR Implementation Guide Set of rules about how FHIR resources are used (or should be used) to solve a particular problem, with associated documentation to support and clarify the usage.

elements of the resource to commonly used standard health domain code systems (e.g. ontologies, classifications) like SNOMED, LOINC or ICD-O.

From a data transportation point of view, differently from previous standards, HL7 FHIR is interoperability paradigm (REST, Document, Messaging, Services) and architecture agnostic, it means that can be used with different system architectures and it allows to access data also at the resource level through REST API. HL7 FHIR API is based on established web standards including XML, JSON, RDF, HTTP, OAuth, and REST. Using these well-understood technologies lowers the barriers to entry and makes it easier and faster to integrate heterogeneous systems.

Finally considering there is wide variability between jurisdictions and across the healthcare ecosystem around practices, requirements, regulations, education, HL7 FHIR also is a "platform specification" providing mechanism to describe, for each usage context, the rules to declare (i) which resource elements are or are not used, and what additional elements are added that are not part of the base specification, (ii) which of HL7 FHIR's RESTful API, messaging and document features are used, and how (iii) which terminologies are used in particular elements.

As described in the HL7 FHIR for FAIR implementation guide[18] HL7 FHIR provides a technical support for implementing the FAIR principles, including I3 "Metadata and data include qualified references to other metadata and data". HL7 FHIR in fact, supports different kinds of (qualified) references among HL7 FHIR resources and to non-FHIR objects. Communities are expected to indicate what are the (qualified) references to other resources that are needed to provide a sufficient contextual knowledge for the scope of their community. This will be formalized by the IDEA4RC project in the IDEA4RC FHIR Implementation Guide.

On top of the HL7 FHIR engine, the OLAP system of the HL7 FHIR capsule will be able to process and extract data in different formats more familiar for data scientists and statisticians to execute their data processing pipelines. The AI analytical models created and trained with federated data processing will be anonymized and shared in a standard format, at this stage of the project one standard under analysis is the open standard for machine learning interoperability Open Neural Network Exchange (ONNX[19]).

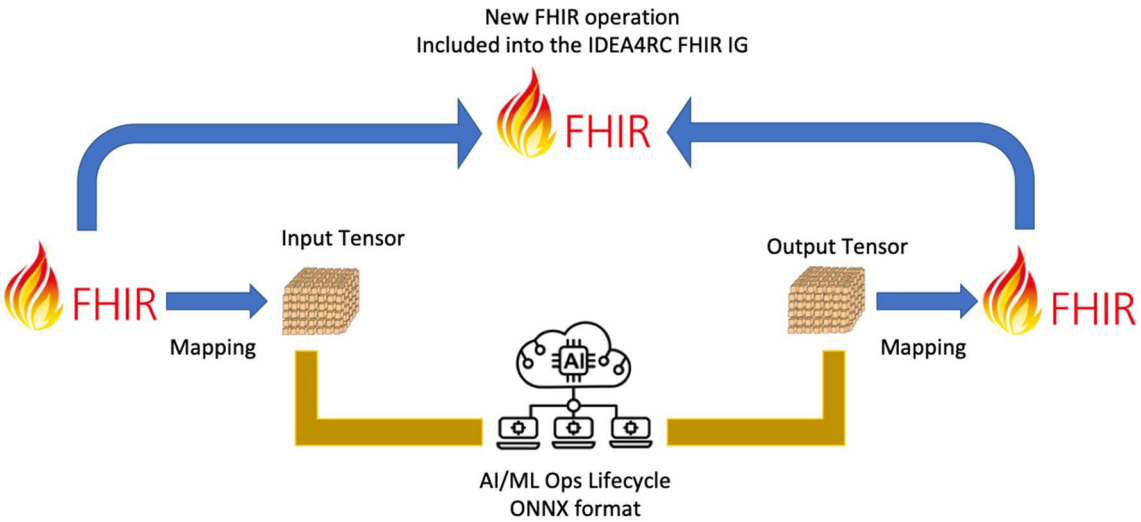

Fig. 4. Interoperability standards for data and AI models in IDEA4RC

[18] http://hl7.org/fhir/uv/fhir-for-fair/

[19] Open Neural Network Exchange , https://onnx.ai/, Last access Dec 2022

Figure 4 shows the relation and integration between HL7 FHIR and ONNX standards. While HL7 FHIR is used for mapping input and output data for an AI service, the ONNX standard is used to standardize the model that uses that input and that output. Once the model is built it will be reinjected into the IDEA4RC HL7 FHIR implementation guide as FHIR operation and shared among all the capsules as a reuse standard policy for federated data analysis.

2.4 Increase the data re-use

According to the FAIR principles[20], "The ultimate goal of FAIR is to optimise the **reuse** of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings". The IDEA4RC platform fully embraces this vision and has the technological and organizational means in place to fully promote reuse of the data involved in the collaboration.

As described earlier in this document, all data from all participating health care organizations will be converted into the FHIR Capsule format. These capsules will contain machine-readable data and meta data, according to international data vocabularies such as SNOMED-CT, ICD-10 and TNM. The harmonization and standardization of data across health care organizations is the foundation of the reusability of the data in IDEA4RC. In order to reproduce results from studies and analyses, a versioning system is maintained within the project. Using this technology, datasets accessed and used for a particular research can be recomposed to ensure scientific soundness

Having federated data with a versioning system, the data can be accessed and analyzed using the federated learning system vantage6. For a study, a researcher gets access to a minimal toolbox of algorithms necessary to conduct the research question at stake. The researcher is invited to publish their analysis script used in publications or reporting. The published script will include versioning of data, vantage6 software and vantage6-compatible algorithms, all required to reproduce the study.

The main objective of the IDEA4RC endeavor is to make data held at EURACAN Centers of Excellence available to third parties, in particular researchers, healthcare managers, and policymakers. Which classes of third parties could be granted access to data, according to what conditions and through what modalities, will be established by Centers themselves (data sovereignty) through the IDEA4RC data governance layer in compliance with current legal and ethical requirements as well as in alignment with the provisions of the future EHDS regulation.

Task T9.3 *Rare Cancer Data Ecosystem Toolkit documentation development*, starting at M28, will produce the documentation intended to support access to the IDEA4RC data ecosystem, for data re-use. The documentation will include a description of the deployed data ecosystem's toolkit components as well as recommendations for reference data use case implementations, that will include lessons learned during the course of the project. Data provenance will be documented and made available as part of metadata and data quality information provided in the frame of the federated data infrastructure established in IDEA4RC.

The metadata regarding data quality will be defined to match the Data Quality Dashboard (DQD[21]) used for OMOP-based systems, but adopted for the context of the IDEA4RC project and FHIR architecture principles. The DQD is based on the Kahn framework[22]. From the description of the Data Quality

[20] "https://www.go-fair.org/fair-principles/"
[21] https://ohdsi.github.io/DataQualityDashboard/
[22] Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw ST, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health

Dashboard: "*Using this framework, the Data Quality Dashboard takes a systematic-based approach to running data quality checks. Instead of writing thousands of individual checks, we use "data quality check types''*.

We will provide a system that can automatically and periodically run data quality checks for properties such as plausibility, conformance and completeness for all the data within the project, analyzing each of those properties in different levels (dataset level, institution level, etc.). Those quality checks can be run in each of the centres locally or in a federated manner if needed and possible.

Record Data. EGEMS (Wash DC). 2016 Sep 11;4(1):1244. doi: 10.13063/2327-9214.1244. PMID: 27713905; PMCID: PMC5051581.

3 OTHER RESEARCH OUTPUTS

IDEA4RC project's beneficiaries will work cooperatively in order to ensure the systematic sharing of knowledge and tools following the principle 'as open as possible as closed as necessary' as described in the "Open Science Approach" Section of the DoA (page 25 Annex 1part B).

The policy for data as well as other research outputs management and sharing is sketched in Figure 4, which includes the different access levels for consortium members and for external users.

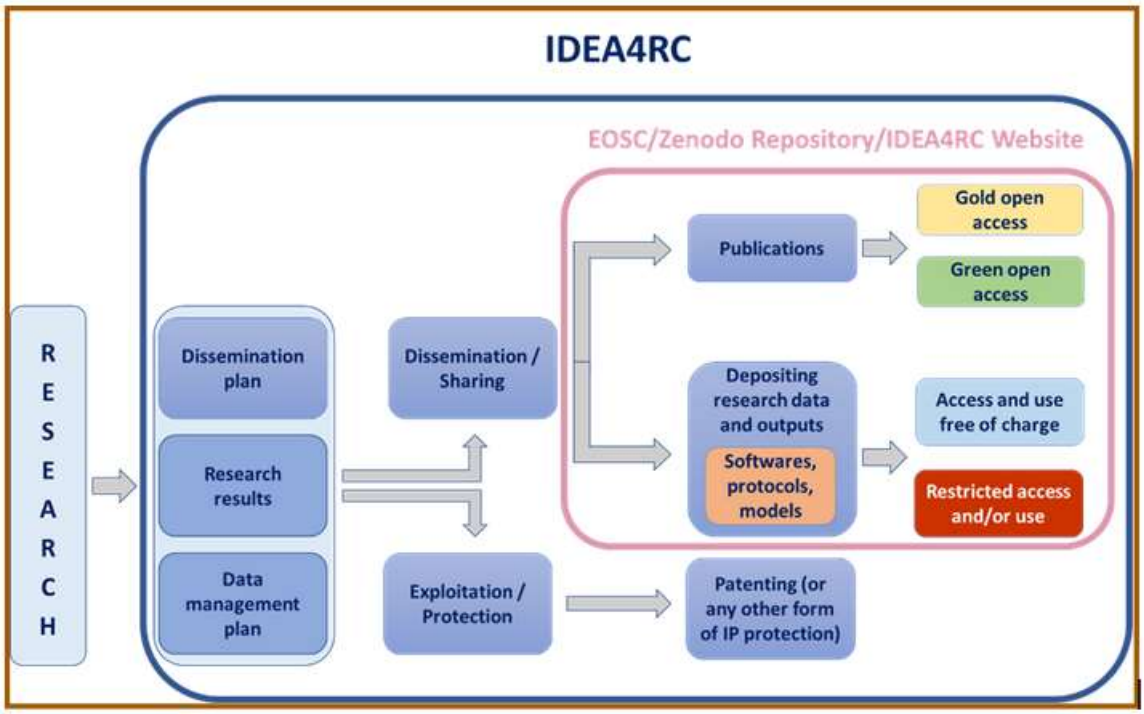

Fig. 5. Data and other research outputs management and sharing

Other research outputs or information that might be generated throughout the project and deemed to be publishable/sharable will also be made openly accessible. However, any dissemination data linked to exploitable results will not be put into the public domain if they compromise their commercialization or have inadequate IP rights and protection.

Moreover, to ensure the FAIRness of data and other research outputs (research software, research publications etc), the project will consider making them accessible via the EOSC catalogue. The EOSC is a federated and open multi-disciplinary environment for hosting and processing research data following the FAIR guiding principles. Since its establishment, the ultimate aim of the "*EOSC is to develop a Web of FAIR Data and related services for science in Europe upon which a wide range of value-added services can be built*". These range from visualisation and analytics to long-term information preservation or the monitoring of the uptake of open science practices.

Depositing any resource (***i.e Services, Catalogues of services, Data Sources*** and ***Research Products***) in the EOSC Catalogue (or Marketplace) does not require their actual upload. Instead, for the way EOSC has been designed, an onboarding process is followed (see Figure 5). According to this, the resource description through metadata is needed along with the link to the original resource and *via* one of the EOSC trusted repositories (e.g. Zenodo, arXiv, eTDR). A full list of repositories is available and searchable on the EOSC website within the **Data Sources** section.

The EOSC Onboarding Process has been designed to ensure that resources provided through the EOSC Catalogue and Marketplace offer the level of quality and interoperability that make them valuable to researchers.

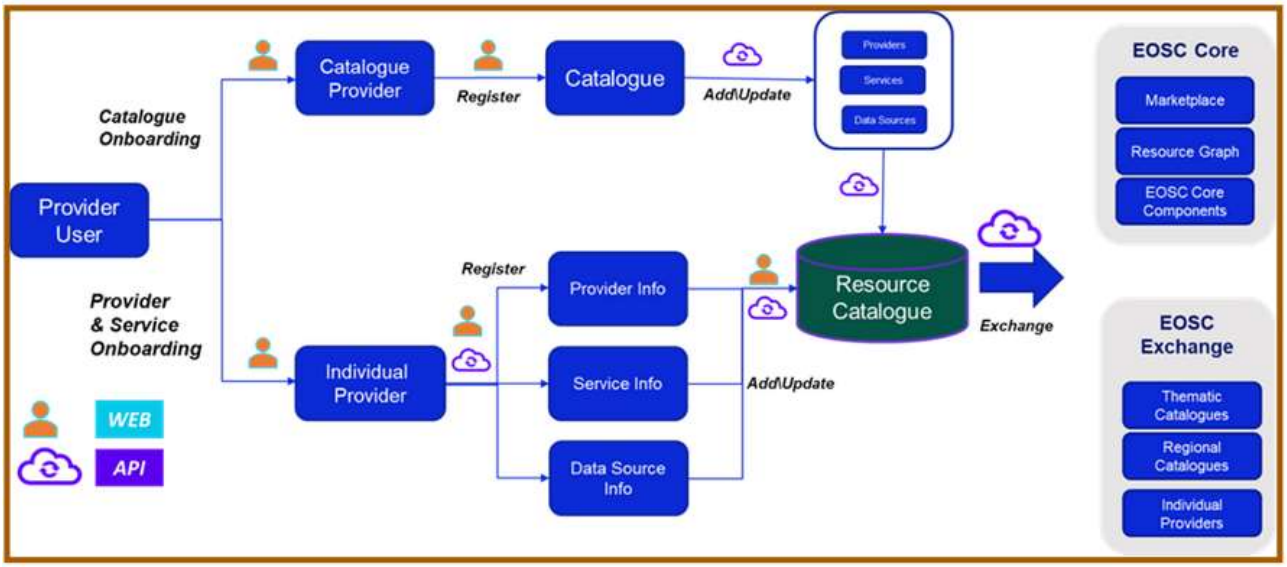

Fig. 6. EOSC Onboarding Process (adapted from https://eosc-portal.eu/)

In addition to this, it is worth to consider that:

- depositing *research products* that are located in (or indirectly referred by) repositories not listed in the trusted EOSC data sources is not permitted.
- depositing *research products* in the EOSC involves costs (see detail in section 4).

After the considerations above are made, each project partner, owner of any type of research outputs (data, software etc..), will be able to decide where to deposit them in the first instance (among one of those trusted and listed in the EOSC data sources), and afterwards consider whether or not to onboard in the EOSC (in Figure 6 a taxonomy about the main ***Research Products*** that can be published in EOSC).

More details will be provided in the next editions of the present document, when the research outputs will be defined and the intentions of the relevant (IPR owners) beneficiaries will be clarified.

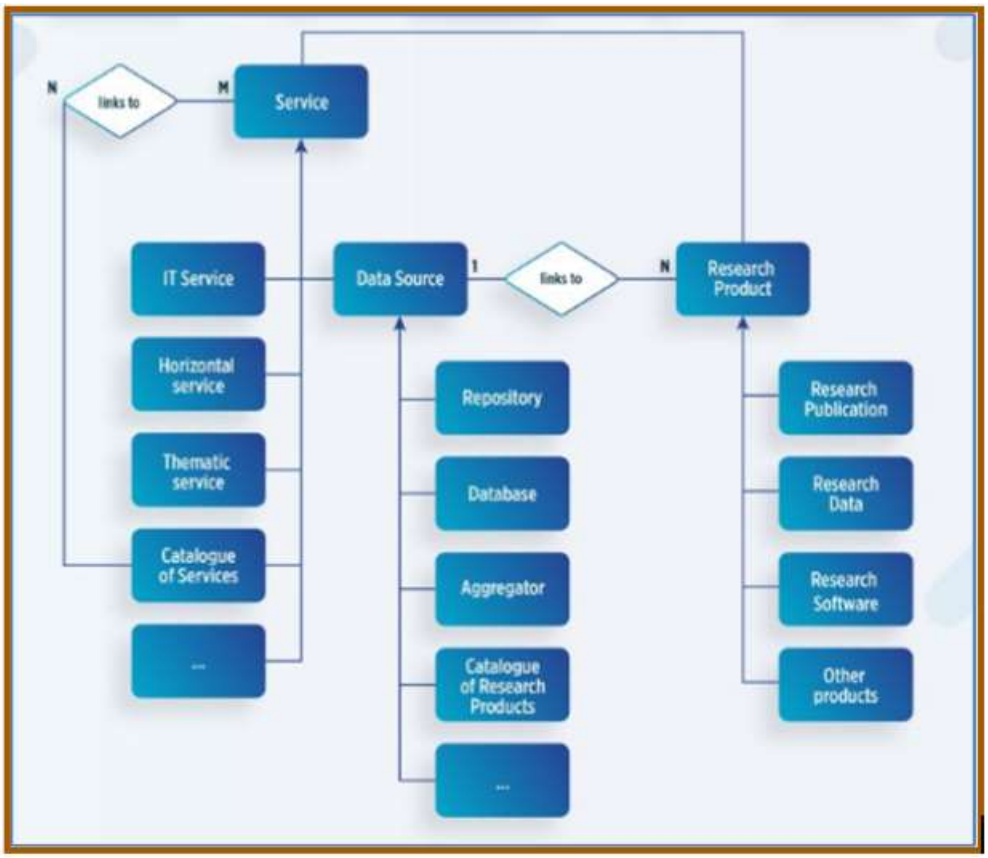

Fig. 7. EOSC information model (adapted from https://eosc-portal.eu/)

In Figure 8, an example of the publishing process as workflow and for different types of *research products* is provided.

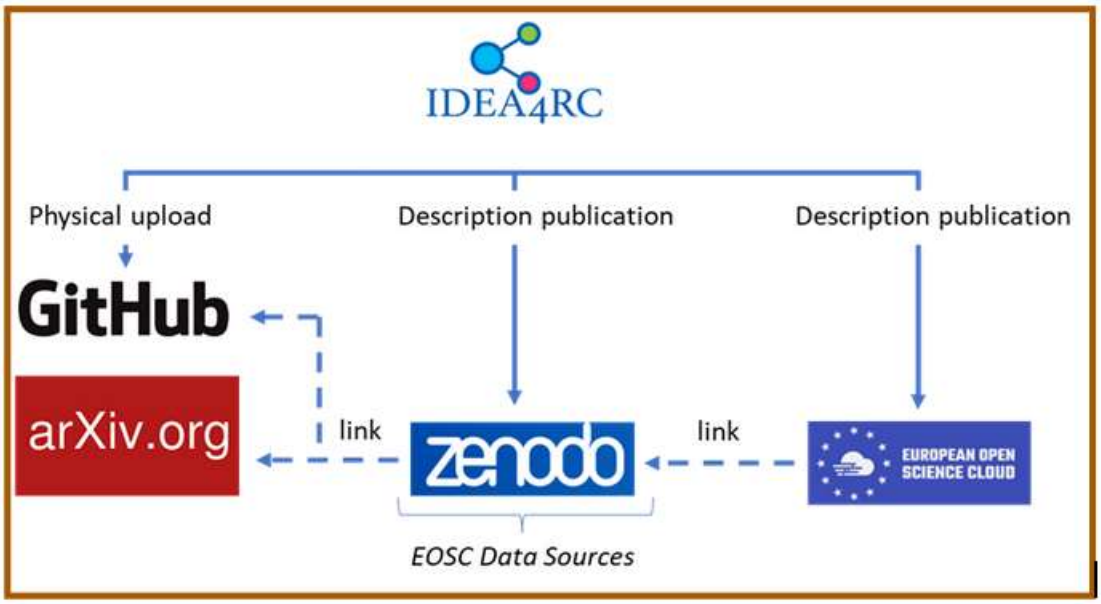

Fig. 8. Example of workflow for the publishing process of different research products

4 ALLOCATION OF RESOURCES

Based on what has been reported in the previous section (sec.3) as well as the DoA document, the costs foreseen to make data and other research outputs FAIR are mainly related to: *i)* physical upload of the research results in a repository (e.g. Github in the case of software, journal repositories in the case of scientific publications) and *ii)* EOSC affiliation.

With regards to the former, costs can vary depending on the type of research result and the repository chosen. For instance, in the case of software for which the owner has decided to adopt an open-source type of release, it is possible to choose a repository such as Github. Whereas, in the case of scientific publications and as described in the Description of Action, the Project results will be published mainly at fee-based open access scientific journals, following the OA Gold method and for which the fee costs can vary broadly. Although, costs related to open access of research data in Horizon Europe are eligible for reimbursement during the duration of the project under the conditions defined in the Grant Agreement.

In relation to the second point, as described in section 3 of this document, it is worth to mention that ***Research Products*** are not directly onboarded to EOSC, but indirectly linked to EOSC through ***Data Sources*** that refer to them and that do not require additional costs (e.g. Zenodo). The following table 1 summarizes the types of costs described above.

Table 1. Example of costs for different types of research products

Research product	Repository/archive costs	Cost Data Source (EOSC)	EOSC affiliation costs
Software	e.g. GIT (different plans[23]): - Free=0 - Team= $48 - Enterprise= $252	(e.g. ZENODO) FREE	Depending on the profile [24] - Member¥ - Observer§
Data	FHIR capsule infrastructure maintenance costs	(e.g. ZENODO) FREE	
Publication	e.g. *Journal of Artificial Intelligence in Medicine* (Gold open access publication fee= $ 3230[25]]	(e.g. ZENODO) FREE	

¥Members: they have a presence in an EU Member State (MS) or Associated Country (AC), or any other country associated with the EU Framework Programme for Research and Innovation.

§Observers: they may be established outside an EU Member State (MS) or Associated Country (AC), or any other country associated with the EU Framework Programme for Research and Innovation.

Therefore, the only costs in order to make them accessible via the EOSC would be those connected to the EOSC affiliation/membership (https://eosc.eu/join-association). All the costs reported in the table above will be covered by the project budget and as assigned to each partner.

[23] https://github.com/pricing

[24] https://eosc.eu/join-association

[25] https://www.elsevier.com/journals/artificial-intelligence-in-medicine/0933-3657/open-access-options

With regards to data management, each beneficiary leading work packages is responsible for preparing the datasets to make FAIR the data collected within its own activities, provided that the project Coordinator (Fondazione IRCCS Istituto Nazionale dei Tumori di Milano) currently coordinates the DMP. The responsibility for general coordination and supervision of the data management will be further discussed as the project progresses.

Regarding the cost of long-term preservation of data, detailed information is quite premature to be given at this stage. Hence, further details on approaches for long-term preservation and accessibility of project research outputs, beyond the end of the funding period, will be provided in future versions of this document and as the project progresses. Moreover, as mentioned before, depositing research products into a repository and make them accessible through the EOSC will contribute to guarantee the long-term preservation of the project research outputs. Ultimately, IDEA4RC will consider to exploit other support infrastructures provided by the EC towards data preservation, e.g., the Horizon Results Platform.

5 DATA SECURITY

In IDEA4RC, data and other digital information will be protected from unauthorized access, corruption or theft throughout their lifecycle. This protection will include every aspect of information security, from the physical security of hardware and storage devices to administrative controls and of access, as well as the logical security of software applications. It will also include organizational policies and procedures.

The security measures enacted by the project consortium and by each individual partner concerned will be further analyzed in the next iterations of the Data Management Plan.

Both FHIR capsules than the federated environment that orchestrate the capsules (the IDEA4RC platform) will be developed with the same technologies and components, the only difference relies on the fact that the capsule manages the data while the federated environment manages the control. The initial defined architecture (at this stage of the project) is similar to the 5G architecture where an access management layer in the Core Network manages users and data in the GNodes network. In IDEA4RC the 5G Core Network is the federated environment while the 5G GNodes network is represented by the FHIR Capsules.

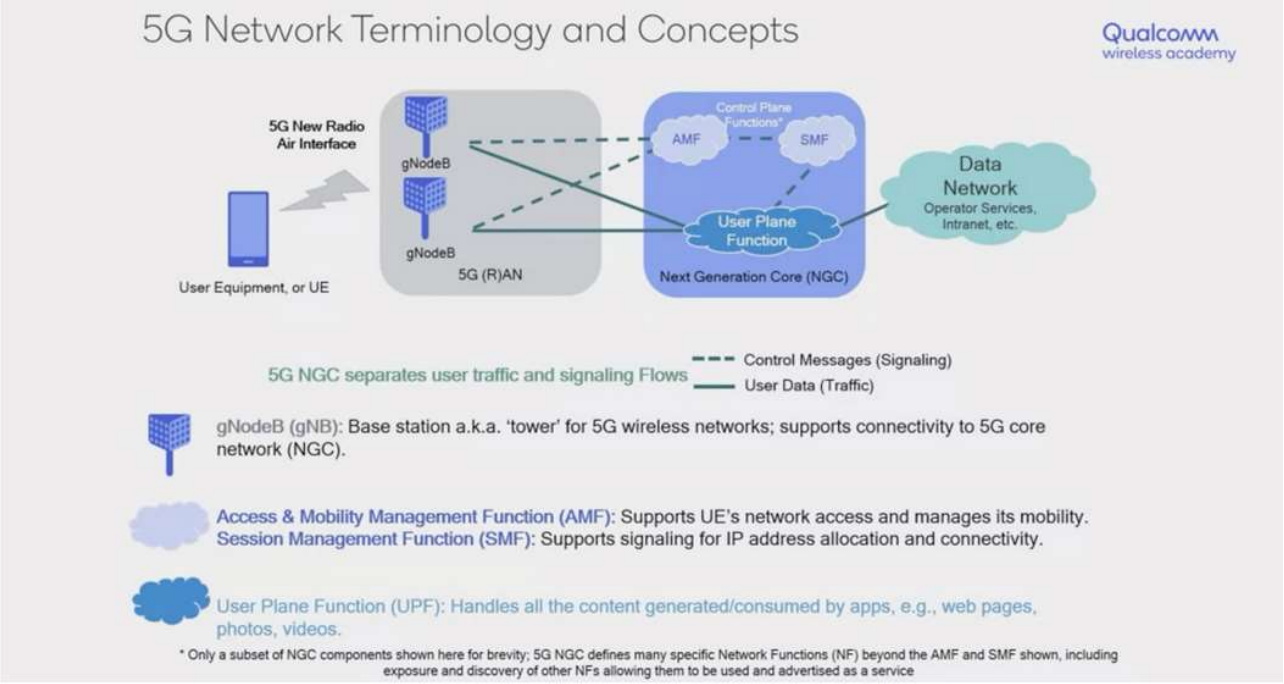

Fig. 9. IDEA4RC relies on 5G Concepts

Both capsule and federated IDEA4RC platform will be based on microservices that will be deployed into dedicated container platforms. Each instance of FHIR capsules as well as the IDEA4RC platform instance are deployed into isolated tenants that are conceptually and physically segregated among each other to grant higher level of decoupling and security.

The FHIR capsules will remain and will be stored and managed within the data providers' premises (hospitals), and, as such the procedures to ensure GDPR-compliant and secure retention and storage of data, data transmission and sharing, personal data pseudonymisation[25] and/or anonymisation, where applicable, as well as general cybersecurity and recovery measures will be applied as foreseen by each Institution.

In order to grant security, privacy and trust for the FHIR capsules and the federated platform several measures at different levels will be applied.

Organizational measures will regulate different levels of access: identification, authentication and authorization. Such measures will be managed and controlled by every capsule administrator and by the administrator of the IDEA4RC tenant for the federated platform. The Governance layer of the federated platform will manage access based on policies and agreement between the data users and the FHIR capsules owners.

Technical measures, in every environment, will be provided in order to: avoid non-authorized accesses, detect threats over both infrastructure and service platform, reduce as much as possible vulnerabilities, avoid data compromise and leakage.

Physical measures will be guaranteed to guard and layered access to the venue space and badge of registered limited users to access the data-room where tenants of the FHIR capsules and the platform are physically located.

The implementation of such measures will cover the following initial identified requirements:

- restrict access to the federated platform and FHIR capsules to authorized users;
- minimize the risk of the unauthorized reading, copying, modification or removal of electronic health data hosted in the FHIR capsules;
- limit the input of electronic health data and the inspection, modification or deletion of electronic health data hosted in the FHIR capsules only to the FHIR capsules owners;
- ensure that users have access only to the electronic health data need for their researches, with a granular access control mechanism;
- keep identifiable logs of access to the FHIR capsules for the period of time necessary to verify and audit all processing operations in that environment;
- ensure compliance and monitor the security measures referred to mitigate potential security threats by means of regular audits of the environments.
- the IDEA4RC platform shall ensure that electronic health data can be uploaded only by FHIR capsule holders or entity that have signed a processing or co-processing agreement with them and can be accessed into the FHIR capsules only with a regulated access (e. g. approved data permit).
- The researchers shall only be able to download non-personal electronic health data from the FHIR capsules.
- Backups and replication of data into the IDEA4RC platform and FHIR capsules shall be maintained secure in every environment.

The procedures used, the methods and applicable timeframes for the review, maintenance and update of security measures will be detailed in the data management and data sharing agreements that will be defined during the project as part of the Data Governance tasks. These procedures will be better detailed in the next releases of the Data Management Plan.

6 ETHICS

No data will be collected from the patient in IDEA4RC. The project will reuse existing data as detailed above. Data reuse ethical framework is provided in the following sections.

6.1 Regulatory framework

IDEA4RC is bound to abide by any ethical and legal requirements applicable to its activities. With regards to the data generated and/or shared during the project, partners are to consider the following general framework described below. Of course, the requirements analysed below are only setting the main framework related to the project's data. A more in-depth analysis will be performed in Deliverable D2.3 Ethical data governance and reuse incentivization approach, ethical and legal analysis and rules to support the ecosystem's "data economy", based on stakeholders' value maximization and trust.

6.1.1 Applicable regulations

General Data Protection Regulation

The GDPR, the main instrument in the EU on the protection of personal data and privacy, will be considered where personal data is to be utilised in the context of IDEA4RC. The project intends to use mainly anonymised data, which is not considered personal data. However, if personal data is to be collected and processed, partners are to abide by the following main principles provided by the GDPR:

1. Lawfulness, in particular considering the sensitive nature of the data that may be processed, health data,
2. Fairness,
3. Purpose, storage and time limitation,
4. Data minimisation,
5. Data protection by default and by design,
6. Accuracy, integrity and confidentiality of data, and
7. Transparency and accountability.

Additionally, data subjects' rights must be respected, while if high risks are identified for the patients, a Data Protection Impact Assessment (hereafter 'DPIA') needs to be performed.

Where personal data is to be shared with third parties, the provisions provided in Chapter V of the GDPR shall be followed. In particular, when data will be shared with third countries or international organisations, they must be done mainly on the basis of one of the options below:

- An adequacy decision issued by the European Commission,
- Binding Corporate Rules,
- Standard Contractual Clauses adopted by the European Commission,
- Approved Codes of Conduct.
- An approved certification mechanism.

Medical Devices Regulation

Given the focus of IDEA4RC on the health sector, the Regulation on Medical Devices may be relevant throughout different phases of the project. Safety and transparency can be found at the centre of the new framework, with most of the provisions focusing on the two.

Indicatively, the Medical Devices Regulation provides for strict ex ante risk assessment procedures in order to ensure that medical devices circulating in the Union are indeed safe for patients. As such, documented quality and risk management plans must be established.

At the same time, an overview of all medical devices is envisioned to be stored in the database of EUDAMED as a means to ensure traceability and transparency.

6.1.2 Upcoming Regulations

European Health Data Space Regulation

The European Health Data Space (hereafter 'EHDS') Regulation is focusing on building the appropriate framework not only for the primary use of data, but also for the secondary use of data, i.e. the use of data for a different reason than the one for which they were originally collected. Therefore, the Regulation is built on the following main requirements that are relevant for the project:

- Easy access to and sharing of data with the aim of improving health care delivery,
- Interoperability and security as mandatory requirements for health data,
- The possibility to use health data for research, innovation, public health, policy-making and regulatory purposes,
- The establishment of a common European format for health data,
- The establishment of new decentralised EU infrastructure for secondary use of data that will support cross-border projects.

Artificial Intelligence Act

IDEA4RC will consider the provisions of the upcoming Artificial Intelligence Act, aiming at regulating the use of AI on the basis of a risk assessment of the system. Due to the importance of the work carried out by the project and the sensitive nature of the analyses that will be performed, it is not impossible that certain activities are identified as high risk. As such, a risk management system may be required aiming at identifying and analysing any foreseeable risks, estimating and evaluating said risks, as well as adopting suitable measures to combat and/or prevent them where possible. Appropriate, relevant, representative, free of errors and complete data governance and management practices must also be adopted.

It is also highly important to ensure traceability, transparency and interpretation of any output produced by the algorithm, while also permitting human oversight. Again, the principles of accuracy, robustness and cybersecurity throughout the AI's lifecycle are highlighted.

Data Act

In line with the EU general policy towards enhancing data use and accessibility, the Data Act provides a number of provisions on data sharing that may be relevant for IDEA4RC. In particular, it reinstates that any personal data sharing must respect data subjects' rights and data protection principles, while a minimum set of information on the data flows must be provided.

Moreover, any data sharing to be carried out must be performed in a fair, reasonable and non-discriminatory manner. Technical and organisational measures must be put in place to ensure data remains safe and accurate while preventing unauthorised access.

Of particular importance to the project are the provisions on interoperability of data, embodying and further specifying the FAIR principles described above.

Data Governance Act

The Data Governance Act is aspiring to become the cornerstone for data use for research purposes. Evidently, the Data Governance Act is of utmost importance to IDEA4RC where the re-use of data held by public bodies will be highly relevant. According to the relevant provisions, a number of conditions must be put in place by public bodies that will allow for the re-use of data in a non-discriminatory, proportionate and objectively justified manner.

Non-profit entities that focus exclusively on the collection of data for reasons of general interest are to have access to the data on the basis of data altruism, after having registered to the relevant record.

6.1.3 Ethics, personal data protection and management

In order to ensure compliance with legal and ethical requirements, IDEA4RC has created a dedicated task (T10.1) that will be in charge of supporting and coordinating the Ethics and personal data protection and management.

In the context of this task, Deliverable D10.1 *Ethics guidelines for enlargement addressing ethic issues arising in the wider community beyond IDEA4RC pilot cases* will focus on identifying and/or adapting the guidelines developed beyond the duration of the project. The development of said guidelines will benefit from the work performed in WP7 to identify the relevant guidelines that can be of use to the entire scientific community.

6.1.4 Data Protection Coordination Board

The Data Protection Coordination Board has been established in the context of the IDEA4RC project in order to assist in and further facilitate compliance with ethical and legal requirements. The DPCB is comprised of the partners' Data Protection Officers (hereafter 'DPO') since they are better equipped to discuss and provide solutions in the challenges that may arise in relation to data processing.

6.1.5 Clarification on the Responsibilities

Coordinator

The project's coordinator is in charge of overviewing the activities of the project. Even though they do not assume responsibility for each partner's legal and ethical compliance, they provide guidelines on how these can be achieved, and assistance whenever required.

Partners

Each partner is directly legally responsible for their data processing activities. Each partner acting as a data Controller is responsible for ensuring the lawfulness of its data processing activities, including any sharing of data with third parties such as partners. Where additional assistance or guidance is required, they may request so by the Task Leader of Ethics, Personal Data Protection and Management or the Data Protection Coordination Board, while maintaining the responsibility of compliance.

Task leader of Ethics, Personal Data Protection and Management

The task leader in charge of data management is bound to support the work performed by the Consortium. Reporting at project level, they are part of the Data Protection Coordination Board. They are in charge of assisting partners when requested and providing additional guidance on compliance matters where that is required.

Data Protection Coordination Board

In order to facilitate compliance with personal data protection provisions, IDEA4RC has implemented a Data Protection Coordination Board (hereafter 'DPCB') that will be in charge of providing guidance and supporting the partners in their activities involving personal data. The DPCB will play an active role in the determination of a data protection strategy and the drafting of a Data Protection Policy. As such, it is also in charge of assisting in the performance of the required DPIA and the drafting of any contractual obligations among the parties with regards to data processing or sharing.

6.1.6 Ethical Requirements

Partners maintain the sole responsibility of complying with ethical requirements. For this reason, most of the partners involved in IDEA4RC already have their own ethical committees within their organisation that will be leading the internal ethical compliance.

As an additional step towards such compliance with ethical requirements, IDEA4RC has also dedicated two tasks, T2.2 and T10.1, on the development of an ethical framework for the operation of the project during its lifecycle but also beyond its duration.

As such, *D2.3 Ethical data governance and reuse incentivization approach, ethical and legal analysis and rules to support the ecosystem's "data economy", based on stakeholders' value maximization and trust* will focus on identifying any ethical risks relevant for the project, as well as proposing an ethical data governance model and guidelines. Similarly, *D10.1 Ethics guidelines for enlargement addressing ethic issues arising in the wider community beyond IDEA4RC pilot cases* will aim at identifying the ethical issues that will be relevant for the enlargement phase, as well as the guidelines that will facilitate it.

6.1.7 IDEA4RC Policy

As was analysed above, IDEA4RC will adopt a clear policy that will differentiate the handling and management of data in two main categories:

A – Personal Data

As explained above, the GDPR will be applied where personal data is involved. Even though anonymised data is no longer considered personal data, this is not true for pseudonymised data and the secondary use of personal data.

According to Article 89 GDPR, personal data may be further processed for research purposes as long as appropriate safeguards are in place to protect data subjects' rights and freedoms. Such safeguards may for instance include pseudonymisation, i.e. "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information". Additional information that may lead to the re-identification of data subjects must be kept separately in a secure environment.

Given that in the most cases, the secondary use of personal data includes patients' health data, it is of high relevance to IDEA4RC. The GDPR leaves room for further specification of the provisions by each Member State, as will be analysed in D2.4.

As explained above, for the secondary use of data, the upcoming EHDS will play an important role on the definition of the project's activities. Until the final adoption of the Draft EHDS, certain changes may be implemented, however it still provides a number of clarifications on the expected requirements for secondary use of medical data and is, thus, already considered in the development of the project's upcoming policies and procedures.

B – Non Personal Data

IDEA4RC is dedicated to maximising access and reusability of non-personal data not only for partners but also third parties that may benefit from such access. Given the importance of the project's activities, it is essential that the datasets generated become available to the scientific community. As a result, interoperability of datasets is crucial to IDEA4RC as it has already been identified that open datasets lead to more accurate results of higher quality.

6.1.8 Data Sharing Policy

As already mentioned, a Data Sharing Policy is going to be established, providing the framework for data sharing among partners and with other parties, based on the work performed during the co-creation workshops envisioned.

Due to its importance for the project, the Data Sharing Policy will be further validated by the DPCB and will be published in full in the context of *D2.3 Ethical data governance and reuse incentivization approach, ethical and legal analysis and rules to support the ecosystem's "data economy", based on stakeholders' value maximization and trust.*

Indicatively, the Data Sharing Policy is intended to be covering:

1. The conditions under which personal data may be shared with partners during and after the project,
2. The conditions under which personal data may be shared with third-parties during and after the project,
3. The conditions under which non-personal data may be shared with partners during and after the project,
4. The conditions under which non-personal data may be shared with third-parties during and after the project,
5. The rules, rights and obligations of partners developing IPR either alone or jointly with other partners,
6. The conditions under which partners may proceed to publications,
7. The rules on the use of and access to data repositories, including potential safeguards that may be in place.

7 OTHER ISSUES

At present the project does not envisage any other procedures for data management.