D8.1 Draft Strategic Agenda

NOTE: THIS DOCUMENT HAS NOT BEEN APPROVED BY THE EUROPEAN COMMISSION AND REFLECTS ONLY THE VIEWS OF THE HEALTHYCLOUD CONSORTIUM

Document Information

Contract Number	965345
Project Website	http://www.healthycloud.eu/
Contractual Deadline	M20, October 2022
Dissemination Level	CO
Nature	R
Authors	S. Canham (ECRIN), C. Ohmann (ECRIN), J. Demotes (ECRIN), M. Panagiotopoulou (ECRIN)
Contributors	M. Matei (ECRIN), N. Blomberg (ELIXIR), S. Scollen (ELIXIR), A. Merchant (ELIXIR), A. Chatterjee (ELIXIR), G. Saunders (EATRIS), J. González Garcia (IACS), E. Bernal Delgado (IACS), S. Garcia Armesto (IACS), L. Portell (BSC), S. Capella Gutierrez (BSC), A. Valencia (BSC), A. Jene (BSC), L. Lopez (BSC), A. Thorogood (UNILU), R. Becker (UNILU), V. Satagopam (UNILU), W. Gu (UNILU), I. Schluender (TMF), P. Holub (BBMRI), M. Th. Mayrhofer (BBMRI), E. Garcia Alvarez (BBMRI), I. Kessissoglou (Sciensano), N. Schutte (Sciensano), S. Cosgrove (Sciensano), A. Martínez García (SAS), C. Alvarez Romero (SAS), C. L. Parra Calderón (SAS), S. Barysch (EMBL), K. Lekadir (UB), T. D'Altri (CRG-EGA), H. Wagener (CHARITE), R. Martinez (IRTIC-Universitat de València), M. Dietrich (EGI), G. Sipos (EGI)
Reviewers	E. Bernal Delgado (IACS), H. Wagener (CHARITE), R. Martinez (IRTIC-Universitat de València)
Keywords	EOSC, EHDS, Health Research and Innovation Cloud, Strategic Agenda

Notice: *The HealthyCloud project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement Nº965345*

Change Log

Version	Author	Date	Description of Change
V0.1	S. Canham (ECRIN)	20/07/2022	Initial Draft
V0.2	S. Canham, C. Ohmann, M. Panagiotopoulou, J. Demotes (ECRIN)	24/07/2022	Internal ECRIN review: Summary table of proposals added, Figure 1 revised, values section expanded, typos corrected, clarification of some paragraphs.
V0.3	HealthyCloud WPs	09/09/2022	Review by the HealthyCloud WPs completed: Several comments and suggestions including changes in Figure 1, addition of descriptions of EOSC and EHDS as Appendix, addition of a "Gaps, Needs and Uncertainties" section, clarifications of concepts in section 1.4, 2 and 3, restructuring of section 3 in the form of "proposed HRIC services", correction of typos.
V0.4	S. Canham, C. Ohmann, M. Panagiotopoulou (ECRIN)	17/10/2022	ECRIN revision based on HealthyCloud feedback completed: Amending according to the feedback received from the HealthyCloud WPs.
V0.5	E. Bernal Delgado (IACS), H. Wagener (CHARITE), R. Martinez (IRTIC-Universitat de València)	25/10/2022	*HealthyCloud internal review process:* Suggestions to align with the terminology used in the EHDS proposed regulation, e.g. Secure Processing Environments, HealthData@EU when explicitly referring to secondary use infrastructure etc. Different suggestions for improvement of the overall text; main focus on "Gaps, Needs and Uncertainties" and "Possible HRIC services".
V1.0	S. Canham, C. Ohmann, M. Panagiotopoulou, J. Demotes (ECRIN)	07/11/2022	ECRIN revision based on the HealthyCloud internal reviewers' feedback: Amending the document according to the feedback received from the HealthyCloud internal reviewers. Proof-reading and delivery to the HealthyCloud coordination.
V1.1	J. González-García (IACS)	17/11/2022	Adjustments for Zenodo publication

			(Final Change Log entries reserved for releases to the EC)

Table of contents

Abstract

The European Commission (EC) organised in 2018 a workshop to explore the possibility of establishing a cloud for health research and innovation, to be accessible by researchers and health professionals throughout Europe. As a direct result of this workshop, the HealthyCloud project was funded with the aim of developing the Strategic Agenda of a Health Research and Innovation Cloud (*in short "HRIC"*) together with relevant stakeholders in the interface of health-related research and cloud technologies. The present document is the **first draft of that Strategic Agenda** and it has been designed to support upcoming stakeholder discussions that will feed into the final HRIC Strategic Agenda, to be provided at the end of the HealthyCloud project.

Section 1 provides an overview of recent developments in the European research, regulatory and infrastructure landscape, with an emphasis given on implications deriving from the establishment of the European Open Science Cloud (EOSC) and the European Health Data Space (EHDS). Potential gaps, needs and uncertainties are identified such as the need to support greater interoperability of health-related data (*e.g. between routinely collected health data and clinical trial data),* the need to provide clear legal guidance to researchers, the need of meaningfully engaging the public into the scientific research processes and orchestrate training, the need to develop technical infrastructure such as the Secure Processing Environments as described in the proposed EHDS regulation, etc.

Section 2 sets the HRIC core vision and values. The HRIC should, among others, be collaborative and promote the application of the FAIR principles while upholding the protection of privacy and promoting the safe use of citizens' health data. Very importantly, the HRIC should be integrated with existing systems where possible and not seek to duplicate services already available (e.g. within the EOSC or EHDS).

Based on the above, **section 3** details 10 specific proposals for services that the HRIC could offer to complement the current European landscape in response to the gaps, needs and uncertainties identified: 1) *A monitoring service for health-related research*, 2) *A legal / regulatory guidance service*, 3) *A training service for researchers*, 4) *A metadata standards service*, 5) *A data interoperability service*, 6) *An EOSC sensitive data users service*, 7) *An "EOSC Health" catalogue service*, 8) *An "EOSC Health" resource service*, 9) *A research community interface service, with HealthData@EU*, 10) *A research community interface service, with the general public*.

Sustainability considerations for the development of the final Strategic Agenda are covered in **section 4**. At this stage, the Strategic Agenda remains *agnostic* about the organisational umbrella or "brand" under which HRIC services are developed (e.g. under EOSC, EHDS or other). This will depend on the availability of funding to support the identified HRIC services. It might be that, as happened with the development of many ESFRIs, further preparatory projects are necessary to explore the implementation of different proposals more deeply. An interesting model for large scale, longer term initiatives within Horizon Europe is provided by *"EU Missions"*, and it may be that this model could be applicable to a HRIC.

The next step in the process of the development of the Strategic Agenda consists of extensive discussion of the proposed HRIC services with different key stakeholders in a

series of workshops starting in December 2022. The stakeholder input will feed into the final Strategic Agenda for the HRIC to be delivered in August 2023.

1 Introduction – The Requirements

1.1 Genesis and purpose of this document

In March 2018, the Health Directorate of the EC's Directorate-General for Research and Innovation (DG RTD) organised a workshop to explore the possibility of establishing a cloud for health research and innovation, to be accessible by researchers and health professionals throughout Europe.

The workshop (reported in [1]) generated **five principal recommendations** for a Health Research and Innovation Cloud (HRIC), made *"to the funding agencies and the actors in the field"* shown, in abridged form, in Table 1 below.

Table 1: Recommendations relating to a HRIC, from the DG RTD workshop, 2018 [1] (Rationales abridged from the original publication)

Recommendation	Rationale
Provide and foster standards, good practices, and guidelines necessary to establish the European HRIC.	The HRIC should be supported by predefined standards, data formats, protocols, and templates. The data standards and guidelines applied in the HRIC should be designed to facilitate interoperability.
Develop and certify the infrastructure and services required for operation of the HRIC.	The HRIC should provide computational infrastructures and services and analytical and visualisation tools to all users as a platform to share knowledge, data, and guidelines.
Enable the HRIC to operate within an ethical and legal framework that is adequate for health systems.	A robust ethical and legal framework has to be developed that defines rules for privacy, security, ownership, access, and usage of data within the HRIC.
Establish a proper environment for the training of a new generation of data and medical scientists.	Education and training of health professionals need to be updated with the HRIC in mind, considering both international standards and practices for data sharing as well as national environments and regulations.
Fund public and private initiatives for the development of the HRIC through EU Framework Programmes (e.g. Horizon 2020 and Horizon Europe).	The EU and its Member States should, together with private investors, develop a coherent, ambitious, and long-term action plan supported by innovative funding mechanisms that consolidate the outcomes from the existing project portfolio into a long-term operational infrastructure.

The stated aim of these proposals was to allow health-related data (generated from both research and clinical practice), research and clinical protocols, software, computational resources, methods, and publications to be more easily identified, more widely accessed and more efficiently reused, allowing a more complete application of the FAIR (Findability, Accessibility, Interoperability, and Reusability) principles [2]. A suite of distributed "cloud" based infrastructures were seen as important technical components of the proposed HRIC, and there was general agreement that the infrastructure should be built upon and integrated with the developing European Open Science Cloud (EOSC).

The HealthyCloud project, a Horizon 2020 Coordination & Support Action (CSA), was designed to build on the recommendations of this workshop and the debate engendered by it, and generate – in particular – "a number of guidelines, recommendations and specifications that will enable distributed health research across Europe in the form of a ready-to-implement roadmap. This roadmap together with the feedback gathered from a broad range of stakeholders will be the basis to produce the final HealthyCloud Strategic Agenda for the HRIC" [3]. The project has a budget of €3 million and, having started in March 2021, will end in August 2023.

This document is the first draft of that Strategic Agenda. As Figure 1 illustrates, it is designed to be fed into the stakeholder workshops, to take place at the end of 2022 / beginning of 2023, to trigger the process of developing the full agenda, due at the end of the project.

Figure 1: Timelines of the development of the Strategic Agenda (designed with Freepik.com)

Its current contents are based on the discussions that have already occurred within HealthyCloud, in particular while developing the "Discussion Paper for the development of the Strategic Agenda for the Health Research and Innovation Cloud" [4], an initial attempt to identify the main challenges likely to be faced by health-related research communities in the near future, especially in regard to data processing and data management, and thus the issues that any HRIC could and probably should be concerned with.

More recent discussions have focused on clarifying the "vision" and "mission" of the HRIC, because as discussed below, there has been a lack of a clear consensus around this, culminating in a half-day face-to-face workshop in Paris at the end of June 2022, as direct preparation for drafting this document. These discussions have been largely coordinated by WP8, which within the HealthyCloud project has the responsibility for developing the Strategic Agenda for the Health Research and Innovation Cloud, with the involvement of the relevant stakeholders.

1.2 The changing landscape

A recurring difficulty with discussing the European HRIC has been the rapid development of the research, regulatory and infrastructure framework in Europe, even over the relatively short time period between 2018 and 2022. Key developments include:

a) The ***European Strategy for data***, published in February 2020 [5], which aims to create a single market for data to support Europe's global competitiveness and data sovereignty. "Common European data spaces" were proposed to ensure that more data becomes available for use, in both a commercial and social context, while keeping the organisations and individuals who generated the data in control. Data driven applications are expected to benefit citizens and businesses in various ways, for example by:
    - improving healthcare
    - creating safer and cleaner transport systems
    - generating new products and services
    - reducing the costs of public services
    - improving sustainability and energy efficiency

b) As part of its European Strategy for data, the Commission's proposal in November 2020 of a Regulation on European data governance (***Data Governance Act***) [6], which was finally published in May 2022 [7]. The Data Governance Act was complemented by the regulation on harmonised rules on fair access to and use of data (***Data Act***) published in February 2022 [8]. While the Data Governance Act strengthens the single market's governance mechanism and establishes a framework to facilitate general and sector-specific data-sharing, the scope of the Data Act concerns the actual rights on the access to and use of data.

c) The Artificial Intelligence Act (***AI Act***) [9], proposed by the European Commission in April 2021, which aims to introduce a common regulatory and legal framework for

artificial intelligence, while taking into account ethical considerations. Its scope encompasses all sectors (except for military), and all types of artificial intelligence.

d) The ***Path to the Digital Decade***, proposed by the Commission in 2021 as a concrete plan to achieve the digital transformation of European society and economy by 2030 [10]. The Path to the Digital Decade is intended to translate the EU's digital ambitions for 2030 into a concrete delivery mechanism. It does that by setting up a governance framework based on an annual cooperation mechanism with Member States to reach the 2030 Digital Decade targets [11] at Union level, in the areas of digital skills, digital infrastructures, digitalisation of businesses and public services. It also aims to identify and implement large-scale digital projects involving the Commission and the Member States.

e) A ***proposal for the European Health Data Space (EHDS) regulation*** [12], published by the Commission in May 2022, which specifically includes provisions for the secondary use of health data for research, innovations and the support of policy making (Chapter IV of the proposed regulation, informally known as EHDS2, now **HealthData@EU**). The European Parliament and the Council are currently discussing the draft legislation

f) ***GAIA-X***, a project developing a federation of data infrastructure and service providers for Europe to ensure European digital sovereignty. Initiated by France and Germany and presented to the general public at the Digital Summit 2019 in Dortmund (Germany), GAIA-X is actively involved in enabling health-related data sharing, and claims to be paving the way for the EHDS [13]. The Health-X dataLOFT [14] platform is being built as part of this initiative.

g) The continuing development of ***the European Open Science Cloud (EOSC)***, especially with regard to numbers of services and the development of mechanisms for supporting onboarding [15]. EOSC is itself, of course, a response to the more general and continuing move towards ***Open Science and FAIR data***, as promoted by funders, publishers and scientific communities themselves [16].

h) The further development of ***National plans for Health Data Hubs*** (e.g. the French Health Data Hub, FinData, Spanish National Health Data Space, Germany's Medical Informatics Initiative) [17-21]. This should be seen in the context of a ***greater recognition of the potential value of real world data (RWD)*** not only as a research and policy making resource, as exemplified by the EHDS proposal, and such initiatives as the EHDEN project [22], but also for supporting regulatory decisions, as with the EMA's DARWIN initiative [23].

i) Growing interest in the use of ***Secure Processing Environments (SPEs)***, also known as Trusted Research Environments (TREs) or data safe havens, to allow controlled access to sensitive data [e.g. 12,24]. Such environments are one example of ***developments in infrastructure and technologies*** to support the management of sensitive data, that include progress in encryption techniques (e.g. homomorphic encryption [25]), the development of federated query techniques [26, 27], progress in natural language processing [28], and the application of machine learning to data mining from large datasets [29, 30].

j) ***Continuing health challenges***, most obviously the COVID-19 pandemic, but also, for instance, microbial resistance to antibiotics, health challenges linked to climate change, and new emerging infections such as Monkeypox [31-33].

k) ***Greater public and political awareness*** of both the promise and the risks involved in using and re-using sensitive data. The pandemic has demonstrated the importance of bringing data together from different sources, but mis-steps with poor transparency of data re-use, including to big tech companies for algorithm development [e.g. 34, 35], and of lax security leading to data breaches [e.g. 36] have increased suspicion, for at least a portion of the public, around the long term management of personal health and research data.

Taken together, these developments clearly represent a complex set of concurrent changes. They do not all target health-related research as their prime concern, but all are likely to impact such research for many years.

1.3 Gaps, Needs and Uncertainties

Considered in isolation almost all the developments described above have been recognised as important steps towards better management of data and thus more efficient and effective secondary use of that data. Considered as a whole, however, this multiplicity of changes has also generated concerns and questions, not least because many of the listed developments are still proposals or in an early stage of development, so that it is not always clear how, and when, they will be fully applied or implemented in practice.

Within HealthyCloud in particular, there has been a considerable debate about the implications of these changes for the HRIC, and for the articulation of the relationship of the HRIC with infrastructures such as EOSC and HealthData@EU (detailed descriptions of the EOSC and the EHDS are provided in the Appendix). This section lists perceived gaps or uncertainties that were identified as potential problems for data use or reuse within health-related research. They are *not* in any order of importance or priority, and in fact many overlap or are interrelated with each other.

Some are existing issues which, it was felt, were not adequately addressed by current initiatives, and so are likely to remain or even worsen in the future, while some are more directly related to the development of the initiatives themselves. Not everyone in HealthyCloud believes that all of these issues will become problematic, but it was

felt useful to itemise them here as an aid to further discussion with the relevant stakeholders.

One way of specifying the HRIC would be by describing suitable services and infrastructures to fill the identified gaps. We have deliberately not done that in this section - which focuses instead only on describing the gaps and needs. How a HRIC could be used to resolve or mitigate these issues is covered in section 3, that describes possible HRIC services.

a) The timeframes of implementation – the need for coordination

Many of the current initiatives and proposals, after the necessary development of systems, trialling and subsequent refinement, will take considerable time to implement fully – an EHDS, for example, is likely to take several years to reach full operational status across Europe. This also means that many of the details of these initiatives, the "final shape" of HealthData@EU for example, or the national data hubs, as well as some aspects of EOSC, and the scope and nature of interactions between these major infrastructures, are unknown at the present time.

The resulting uncertainty is a common component of many of the other concerns listed in this section. There is also the worry that different initiatives will proceed in parallel and in relative independence, unable to easily coordinate with other initiatives because all are still in development and "following their own path". For instance, the European Commission has recently reported that "13 Member States have started to put forward more centralised national systems to provide access to health data, but there is still no link between them at EU level, the system remains fragmented and there are differences between tasks, even though they share many commonalities" [37]. The EHDS offers a mechanism to reduce this confusion, and this issue is currently being considered within the TEHDAS JA and the EHDS2 pilot projects. The problem is that the national data hubs are likely to be running, with their own established procedures, before this is possible, making it much harder to then standardise practice, requirements and data structures.

In this context, there would be value in a mechanism that could provide *ongoing* coordination between initiatives. Working broadly across health-related research, beyond the specific liaison already envisaged between some aspects of EOSC and EHDS, such an initiative could monitor progress, identify potential problems, and also access and summarise the views of the "users" – the research communities in particular – feeding back those views to funders and initiatives alike. It could make new services related to data sharing both quicker to develop and more effective once developed. In an era when zoonoses and other major health threats appear to be becoming more common, speeding up the development of current initiatives, having them ready to face the next pandemic, could be of paramount importance.

b) Specifics of sensitive data – the need for sensitive data management support

Sharing sensitive data requires much more time, work and money – consideration of consent and legal issues, de-identification, re-analysis of de-identified data, data use agreements etc. – than sharing of "normal" non-sensitive scientific data. One concern is that the push towards data sharing, particularly in the research sector, will not be

adequately resourced at the level of individual studies. The risk is of developing infrastructures and systems for storing and sharing data, but then, because of the difficulties and costs of preparing data for secondary use, having insufficient material to fill them.

Globally there have been some efforts to promote data sharing prospectively. In the U.S.A. the NIH, for example, as of 25 January 2023, will require that all scientific data from studies it funds are shared in a timely and responsible manner, and the data sharing plans are a mandated part of the funding application [38]. Similar large-scale initiatives in Europe, however, have yet to appear. Currently, data sharing of sensitive data is not normally funded in EC projects, e.g. no budget is allocated for de-identification of data or long-term storage in repositories beyond the end of the project (although funding has been made available for increased use of RWD, e.g. in the EHDEN project). The need for funders to better incentivise and valorise data sharing in health-related research remains a powerful barrier to data sharing and re-use.

The topic of sensitive data has been gaining importance in EOSC, and services for sensitive data, supported by work in various EU projects, are starting to appear in the EOSC Marketplace, with more to come (e.g., B2SHARE, the repository for clinical research individual participant data being developed by ECRIN and the University of Oslo [39]). In addition, a new Call is soon to be announced, dealing with trusted environments for sensitive data management in the EOSC (HORIZON-INFRA-2023-EOSC-01-06). This is all encouraging but the EOSC still primarily assumes open data, and most of the systems currently being built do not include "access control by design". Protecting sensitive data and controlling access to it is not something that can be easily achieved by simply bolting on additional features to existing systems -- it requires a more fundamental change to the data model underlying EOSC, itself based upon a good understanding of the requirements of the different research communities with regard to managing sensitive data. An example is the need for timely and demonstrable deletion of sensitive data, when it is no longer required at the “point of compute”, for instance an SPE. This introduces additional management and cost requirements into the data life cycle, which must be recognised and integrated within funding processes.

EOSC also assumes public metadata, for example in constructing catalogues of resources. A potentially difficult issue may arise if the catalogues themselves have sensitive metadata that should not be shared openly, e.g. metadata relating to early phase clinical trials that could be commercially sensitive. Handling sensitive metadata may therefore become an additional issue to be tackled within the data sharing process.

c) A rapidly evolving and complex legal landscape – the need for guidance

The GDPR has not clarified all data sharing issues in health-related research. With the derogation available with respect to research data, different Member States can, and do, make use of their own national legislation and regulations. In addition, even legislation with an explicit EU/EEA scope may be interpreted differently by different Data Protection Authorities, or in some cases not interpreted explicitly at all, which can cause greater confusion. Debate still exists about some critical terms - for example the

definition, in practical terms, of “anonymisation”. Consequently, establishing a clear legal basis for the secondary use of data of various types, whether using consent or some other justification (e.g. public interest, legitimate interest), can be difficult, especially across national or even institutional boundaries.

The EHDS, while proposing a new legal framework, explicitly makes the point that it will not amend the GDPR. The question of the interaction between the GDPR and the proposed EHDS legislation, therefore, remains open, and indeed has been labelled “the elephant in the room” [40]. Researchers involved in sharing sensitive data will urgently need this issue to be clarified.

In addition, new European legislation not only takes some time to be enacted, it also often requires a lengthy period of testing in court, before its interpretation becomes more consistent and better understood across different legal jurisdictions. Further time is then needed before it is referenced within relevant regulations and working practices.

The changing legal landscape, including the Data Governance Act and the AI Act as well as the proposed EHDS regulation, will only exacerbate the existing need to develop and distribute clear guidance, policies and descriptions of good practice that health researchers can use, especially those preparing data for long-term storage and potential data sharing, or those seeking to locate relevant data. Such guidance needs to be authoritative, ideally established by discussion between representatives of regulatory bodies (e.g. the European Data Protection Board) and the scientific communities involved.

The ultimate goal should be a clear and homogeneous legislative and regulatory framework for all aspects of health-related research. That may not be achievable, but the process of working with authorities to provide effective guidance, at least in terms of data re-use, should help promote this goal, and as a minimum it should provide a map of how the relevant legislation is applied and interpreted in different areas.

d) Data access organisation – the need to support multinational research

A concern expressed by some has been that, while the EHDS, and developing catalogues within EOSC, will certainly facilitate access to health data – the support for managing and preparing such data for research purposes may not always be as well developed as it could be. The EHDS, for example, should improve access by establishing a health-specific data sharing framework with clear rules, common standards and practices, infrastructures and a governance framework for the secondary use of health data (see Appendix 7.2). It will also provide SPEs where that is deemed necessary by data providers, and proposes national entities (e.g. Health Data Access Bodies - Art.36, National Contact Points for Secondary Use of Health Data - Art.52) to manage access to data for secondary use.

Health-related research (whether, for example, clinical trials, observational studies, or epidemiological work) often makes use of data from several different countries at the same time. It is not clear how this will be facilitated by the proposed HealthData@EU structures, and what additional systems might be required to support data sharing and secondary use in the context of multinational clinical research. The EHDS proposal

addresses this issue in broad terms, but the details of what will eventually be agreed and developed is still to be worked out. It is not clear, for example, if, when and how data that would normally be accessible in different SPEs, in different Member States, might be able to be aggregated within a single SPE. The legislative proposal will enforce a number of things, some others will be developed by the countries as implementation and delegation acts, and finally some others will remain as recommendations, but the final framework is still under construction. In addition, as described below, the data will often be complex in form and extremely heterogeneous in content, leading to very significant challenges in making the data interoperable.

e) Data heterogeneity – the need to support greater interoperability

There is, potentially, a huge range of data types that will be accessible within the developing ecosystem for health-related data. They will have a diverse range of sources (clinical observations, biochemical and biomarker tests, genetic and gene expression data, radiology images, claims data, social science surveys, environmental correlates, etc.), and will have been gathered in different contexts (routine healthcare, interventional clinical research, basic biological research, longitudinal population studies, etc.). Textual components are likely to be expressed in different languages, in a largely unstructured form, while numeric components may be in different units.

Even data of a similar basic type may be structured differently and use different vocabulary / ontology systems. EHR systems, for example, can be very specific to health systems, sometimes to individual clinics and hospitals. It is true that there are initiatives within the EHDS promoting more uniform EHR systems and data (TEHDAS JA is expected to provide further guidance at the end of 2022), and slowly increasing use of common data models such as OMOP, but it will probably be many years, perhaps decades, before EHR standardisation is widespread. In addition, whilst still dealing with clinical observations and laboratory and pathology data, clinical trial data is often implemented using CDISC standards and vocabularies. The controlled vocabulary systems used in routine healthcare and EHR systems (e.g. SNOMED CT) do not easily map to the CDISC based vocabularies more common within clinical research, particularly in industry, and neither map well to those used within social sciences (e.g. DDI) [41]. The need for semantic and syntactic interoperability is widely recognised, but the practical difficulties are not to be underestimated.

There is therefore a need to discuss these issues, before the various data standards' camps become too entrenched, and try to develop systems that can render data more interoperable, and thus more available for re-use (or better still work towards common vocabularies and data structures at source). SPEs will include facilities for analysis but may not have extensive systems for pre-processing and re-organising datasets. Improving the interoperability of the data they contain will therefore often need to be considered before data is moved into such environments.

This is clearly a huge issue that will be difficult to tackle but one that is of fundamental importance. Improving access to data is good, but provides much less added value than providing access to data that uses interoperable data standards, and so is relatively straightforward to understand, and easier (and cheaper) to aggregate or compare with other datasets covering the same data points.

f) Improving findability – the need for better discovery metadata

As the pandemic has demonstrated, there is an increasing need for scientists to be able to identify possibly relevant research, and data, from outside their own immediate domain, to facilitate cross-disciplinary investigations. Currently this can be difficult, because not all potentially valuable data is listed in a catalogue, and researchers may need to "go fishing" for data, using personal contacts or conference networking, and then negotiate access.

There is a clear need for metadata catalogues that are much more standardised than they are now, and that can be applied across the different health research domains. Work is ongoing to tackle these issues (e.g. within current projects such as EOSC-Life [42], BY-COVID [43], EOSC Future [44], FAIRCORE4EOSC [45], and PHIRI [46] and also within some national initiatives, such as NUM-CODEX, spun out from the Medical Informatics Initiative in Germany [47]). EHDS also plans to develop a standardised common European metadata standard to describe available data collections and improve their findability, though some of the specific features of the system remain to be clarified (see Appendix 7.2).

If discoverability is to span all health-related research, however, these and similar initiatives should be encouraged, monitored, and ideally orchestrated, across all the various biological, clinical, environmental and social sciences that are relevant to health. There is a need, in particular, for the different approaches and schemas used for creating FAIR data in EOSC and EHDS to be understood, coordinated and harmonised. Ultimately, the same schema for supporting findability needs to be applied across all health-related research domains.

g) EOSC and EHDS – the need for clarity

Currently, there appears to be some uncertainty about how and where health-related research and secondary analysis of health data could and should be represented within the major infrastructures, involving both EOSC and EHDS. "EOSC-Health" has been proposed as an extension of EOSC, dealing with health services research, epidemiology, public health, health policy, and clinical research among others. EHDS claims much the same set of activities, and the proposed regulation explicitly identifies a wide range of health and health-related scientific data as falling within its remit.

The difficulty is that at the moment, while the HealthData@EU proposals include various forms of research data, many see the focus of HealthData@EU being very much on data obtained from routine healthcare – either directly from EHRs or indirectly from registries or claims data – rather than from clinical, genetic or social science research. The perception is that the sheer scale of the task facing HealthData@EU, in terms of managing RWD obtained from healthcare, will inevitably mean insufficient "bandwidth" being available to easily handle data from other sources.

In due course, the exact scope of both the EOSC and the EHDS will become clearer. Some degree of overlap seems likely, which means that liaison between the two organisations will remain important – perhaps directly, perhaps via an intermediate body. Ideally, communities that make use of both services should also be involved in

that liaison. But some gaps may also remain after the activities of both organisations are considered. These too need to be identified and, if possible, funding obtained to create the necessary missing services.

h) Infrastructure components – the need for SPE development

One specific area of likely interaction between EOSC and EHDS involves Secure Processing Environments. These are foreseen as an important infrastructure component for data management and analysis related to the secondary use of sensitive data, though current provision of SPEs is limited.

In HealthData@EU it is proposed that the Health Data Access Bodies shall provide access to electronic health data only through a SPE, with appropriate technical and security measures in place (guidance is currently in preparation from TEHDAS JA). EOSC could also include a network of SPEs, with clearly exposed governance and security attributes that would allow users to find the best one for their needs, as well as secure communications facilities to allow sensitive data to be moved safely to the right locations for processing.

Within HealthyCloud, the technical aspects of SPEs are being explored by looking at existing national computing infrastructures, analysing the solutions that are being implemented and developing guidelines on how a similar but federated compute infrastructure could be set up across Europe. There is a need to update, complete and implement these guidelines and to use them as input into SPE development.

As SPEs become more common, in both EOSC and EHDS, it will be important to continue to monitor their development and features, perhaps pooling resources and designs to improve their effectiveness, for example in developing and disseminating federated query techniques. It will also be important to incorporate ongoing advances in technology and systems, e.g. in cryptography, including those offered by commercial organisations. The underlying need, however, will be for the continuous identification of gaps and problems in the services provided, especially from the viewpoint of researchers, and then working to close those gaps.

i) Public engagement – the need to retain trust

The pandemic has demonstrated the sometimes uneasy relationship between research, public policy that claims to be based on research, and the general public's beliefs and behaviour. As exemplified by vaccine hesitancy, there is a problem around trust in research and researchers, within at least a section of the public, sometimes based on misunderstanding or misinterpretation, sometimes based on more fundamental social and political beliefs, but partly also based on knowledge of the very real mis-steps and poor transparency that have sometimes occurred.

From both an ethical and pragmatic perspective, it is important - indeed many would see it as an obligation - for researchers to explain the FAIR use of data so that everyone involved in the creation of health-related datasets, including in particular the citizens from whom the data is derived, can understand why reusing health-related data is seen as important, and how FAIRness is promoted whilst still trying to preserve the privacy of individuals. Other salient topics that could usefully be considered include:

- how best to explain, apply and manage consent, including opting-in and opting out systems;
- how to operationalise the return of research results to the community;
- how data altruism may be harnessed and better recognised within health-related research.

If not actively addressed this issue, of increasing and retaining public understanding and trust, risks jeopardising the entire move towards greater use of personal data for research purposes. Exactly how it should be tackled is a more difficult question, but this is a problem that relates to the whole of health-related research rather than individual disciplines, and would be better tackled at that level.

j) Links to research infrastructures – the need for communication mechanisms

Many of the concerns listed in this section have stressed the need for the views and needs of the scientific communities involved in health-related research to be collected, transmitted to the various organisations and infrastructures being developed, and heard –for example to ensure the requirements around sensitive data management are understood, or multinational data access is facilitated, or metadata schemas are developed that meet the requirements of researchers. Research infrastructures (RIs), with existing strong links into the communities they serve, could offer a mechanism for such communication, but there then needs to be a framework for linking those RIs to the major infrastructures – the EHDS and EOSC in particular.

In EHDS, Member States will cooperate at EU level on cross-border digital infrastructures to enable data sharing for secondary uses of health data. National Contact Points will be the entry point into the EHDS, with data holders (primarily Health Data Access Bodies) making health data available for research and policy making. Some EU research infrastructures (e.g. ESFRI life science RIs) may have a role as a user of the EHDS data or producer of data and services to the HealthData@EU community, but in general their objectives and tasks are much broader than this. Many are involved in developing and providing research services, especially cross-border services, in standard creation and quality assurance activities, and in the support of translational processes and in innovation transfer. In the current conception of the EHDS it is unclear how the link between these infrastructures and EHDS will be established and how a continuous exchange of information and views will be achieved while taking into consideration the very different governance models of the relevant ESFRIs.

In the case of EOSC, the EOSC Association actively involves different key stakeholders through specific task forces contributing to the EOSC development. The members of the task forces include different representatives of organisations, projects and initiatives fully committed to supporting the EOSC “vision”. This approach is certainly useful, but it does not guarantee adequate involvement of ESFRIs in the development of the EOSC. In addition, EOSC does not include all the relevant research communities, e.g. the population health research community is not included.

For both EHDS and EOSC, therefore, some mechanism(s) to involve health-related science ESFRIs, most easily collectively through an intermediate forum, could be of benefit to all concerned. It would allow the infrastructures to more easily keep in

touch with the "important issues" of the day, and it would allow existing initiatives to be identified and better coordinated with both EOSC and EHDS, in some cases to avoid "reinventing the wheel", in others as a means to bring in the experience of key stakeholders and explore what succeeded and what did not.

k) Ensuring new services are used - the need for training

The scale and range of developments discussed above reinforce the need for training for researchers, so that they are both aware of relevant resources and can make effective use of them. Most researchers are likely to need some training to make the most effective use of SPEs, for example, and to understand or apply new metadata schemas, or to apply for data under controlled access in a national node. Of course, where possible, systems and tools should be intuitive or familiar enough to be used without excessive training, but that is only likely to be possible with a fraction of the proposed new developments.

Training on specific aspects of health research is already provided to a large extent by RIs and within large projects. There is no need to replace this training - rather the aim should be to complement and to coordinate existing training initiatives with the overall intention of capacity building across Europe, including to facilitate the adoption and use of new data sharing services. Training could also include certification and provision of credentials or a registry for suitably trained researchers.

1.4 The HRIC as an Interface

In summary, as the long list of needs in the preceding section makes clear, the rapidly changing landscape in health-related research is not perceived as having changed, in any fundamental way, the need for some form of HRIC, as identified at the original workshop, and as re-identified within both the HealthyCloud discussion paper and the preceding section.

There is still a requirement for better standardisation and "FAIRification" of data, for services specifically tailored to storing and processing sensitive data, for mechanisms to efficiently integrate real world health data with multi-national health-related research, for clarification of the regulatory framework, for training of researchers and for adequate resourcing of those and related initiatives. If anything, the more recent proposals, such as those for HealthData@EU, may have accentuated these needs, especially in the medium term as proposals slowly evolve into reality.

Recent changes have also brought to the fore questions about the particular "umbrella" (or umbrellas) under which development of these services and initiatives should take place. This will ultimately be determined by funding decisions, and the availability of the relevant Calls, or perhaps even the establishment of some form of permanent infrastructure (not necessarily as a separate organisation, it could be embedded within one or more existing infrastructures). This document makes some suggestions about this, but gaining clarity about the views of major funders around these issues – in particular the European Commission – sooner rather than later would be an important step forward.

To help to move forward in the meantime, the suggestion here is that the HRIC is best considered at the moment as an "interface" of services, a specified list of resources and related functionalities designed to meet identified needs. Users and organisations within the research ecosystem that make use of those services do not need to know the details of how the functionality is resourced, and the organisational infrastructure behind the resource should be of little importance. Instead, they only need to know what services are available, how they can use them, the inputs required and the expected outputs.

Section 3 takes this approach explicitly, and attempts to explore the different elements that might make up a "HRIC interface", focusing on the set of services and resources that could be developed to support better handling of data within health-related research, and to help meet the perceived needs of the relevant research communities.

While any identified set of services will obviously have to have a concrete implementation at some stage, the current HRIC proposals do not – indeed cannot – make any specific statements about implementation, beyond general suggestions to funders (see section 4 on Sustainability). The greater the perceived impact of a particular service on research, and thus, ultimately, on the health of citizens, the greater the number of stakeholders likely to support any particular suggestion and the greater the chance it will be implemented. An essential part of the process of turning this draft Strategic Agenda into the final document will therefore be the clarification and organisation of stakeholder opinion behind specific suggestions for services.

2 Core vision and values

This section attempts to set out a basic framework for the HRIC, applicable both to the HRIC conceptualised as an interface of services and to any concrete implementation of those services.

As an initial step, a core "vision statement" was constructed, given below.

> *The vision of the HRIC is to improve the health and well-being of people by increasing the quality and impact of data-driven health-related research. Its focus is on the most effective and efficient re-use, for research purposes, of health-related data from a wide variety of sources.*

This statement was deliberately kept short and simple, to emphasise the central purpose of the HRIC. It does, however, raise the question of what is meant by "health-related research", a phrase which is used, with minor variations, throughout this document. The phrase was chosen to deliberately be as inclusive as possible, reflecting the huge range of factors that can influence health, and to make it clear that the HRIC should be able to encompass data from a wide range of data sources, to allow the study of any health determinant.

A HRIC would certainly include or reference traditional types of clinical research data, and in addition include mechanisms to import or access healthcare derived data, including that from Electronic Health Records, registries, dispensing records and insurance claims. It should also, however, encompass data from basic biological

research, for example from genetics and microbiology, and from social, psychological, economic and environmental research, where they have relevance for human health. One of the aims (and added value) of establishing the HRIC should be to reduce the barriers between the different domains and data sources involved in health-related research.

A high proportion, though certainly not all, of this data will be sourced from identifiable individuals – it will therefore be personal data whose management will fall under the GDPR and other data protection legislation, and in most cases will also be sensitive data as defined by that legislation. This is the main differentiator of the services required by a HRIC to those of other domains within EOSC, and more similar to those proposed for managing healthcare data under the HealthData@EU.

The attributes listed below were also identified as **key values** or aspects of any HRIC. They are predicated on the assumption that any HRIC will be a relatively persistent initiative, existing beyond the lifetime of a single project.

a) *The HRIC must be collaborative:* Whether or not a HRIC is a sub-part of some other organisation it will need to collaborate with other related initiatives. Much (but not all, see below) of this collaboration will be at the European level, working with EOSC and/or EHDS and their various component parts, or with directorates within the EC, or specific IHI or EU-funded projects, infrastructures and RI clusters. Such collaboration is not an optional addition to the HRIC's activities – it is an essential part of what a HRIC must do, or be, in order to deliver the services the research community needs.

b) *The HRIC must be global, but retain a European focus:* The HRIC will be funded by European funds / money / resources and – especially if embedded within European infrastructures – be used mainly by researchers based in Europe, with a focus on data relevant to the health of European populations. But of course, research is global in nature, and research work and tools (e.g. evolving data standards to support interoperability) are also very often global in scope. The HRIC therefore needs to be open to developments, interactions, and collaborations in health-related research on a global as well as a European level, and plan and implement activities accordingly. Some tools and services, for example, should be built and made available as global resources, just as European researchers currently use resources developed elsewhere (e.g. the Maelstrom repository of cohort research [48], which originated in Canada). This is one way in which the HRIC can help to ensure that the value of European research, and European researchers, continues to be recognised at a global level.

c) *The HRIC must be demonstrably useful to research communities:* The purpose and value of a HRIC needs to be clear to, and supported by the associated research communities, i.e. all those concerned with health-related research. This may sound obvious, but it does mean having clear long-term and short-term aims, and metrics with which to measure them, plus arrangements for ongoing dialogue with the health research community – the HRIC's users – as well as funders.

d) *The HRIC must promote the application of the FAIR principles to data:* This is a central challenge within health-related research and "FAIRness" must therefore be a key driver and metric within all HRIC activities, and apply to all stages of the data life cycle. A HRIC should also contribute to the further development of ideas around what "FAIR" means in practice for health-related research data, especially sensitive personal data.

e) *A HRIC must uphold the protection of privacy and promote the safe use of citizens' health data:* A commitment to FAIR data should never mean that the privacy of individuals is put at risk. At both a technical and policy level, a HRIC should strive to monitor, uphold and promote privacy protection, both within its own systems and within those of referenced data sources. This would be part of a wider task, of ensuring the appropriate and safe use of data by researchers.

f) *The HRIC must be integrated with existing systems where possible:* As an example, if some parts of an HRIC require Authentication and Authorization Infrastructure (AAI) functionality, existing systems (e.g. the Life Science AAI already developed in the EOSC-Life project) should be used rather than a new service developed. The HRIC should use what is already available whenever possible. The same is true of other core EOSC capability, such as monitoring the availability of onboarded services, where they are relevant.

g) Whether or not a HRIC requires a separate web portal remains an open question. If EOSC / HealthData@EU find they cannot easily provide or integrate portal facilities for a high proportion of health-related research resources within a reasonable time frame, then a separate 'HRIC portal' is likely to be necessary. In general, however, users would probably find it easier to use resources and resource descriptors onboarded to, and embedded in, wider EOSC / HealthData@EU systems.

h) *A HRIC must not try to duplicate other existing services:* For example, the management of healthcare data for potential secondary use will largely be handled by entities within the HealthData@EU. There is no point in a HRIC trying to duplicate that functionality, though it might conceivably be involved in complementing the EHDS services by, for example:
   i) Interacting with EHDS to ensure healthcare data resources (often pre-processed in some way to allow their secondary use) are easily visible to researchers, to allow assessment of their potential value.
   ii) Interacting with EHDS to establish systems allowing researchers to more easily import or access that data, especially when comparing or integrating it with other data related resources, or when seeking to use data from multiple countries at the same time.

3 Possible HRIC services

This section contains 10 specific suggestions concerning the services that a HRIC *could* offer. All of them have been suggested at one time or another as potential features of

a HRIC. Many clearly correspond to the gaps and needs listed in section 1.3, although the correspondence is not exact and the ordering of suggestions is quite different. Issues around how such services might be organised and funded are discussed within section 4 on Sustainability.

In all cases it is suggested that the task for stakeholders will be to consider, for each service:

a) Whether it would add sufficient value to justify its cost, most often in the first instance to researchers working with health-related data but, ultimately, as something that can contribute to improving the health and well being of people.
b) Whether the service is already provided (or likely to be provided) by other existing or likely developments, or whether it, (or parts of it), represents a genuine gap that needs to be filled in some way.
c) If not already provided, whether funding for the suggested service is likely to be available, taking into account the various options, both for implementation and sustainability, that might be possible.
d) Whether currently covered elsewhere or not, whether the activity should be included within a "HRIC", and "branded" as such, or is more easily developed within existing or proposed structures (for example EOSC or EHDS).

Whilst future funding decisions remain unknown, the purpose of this section is to help the community arrive at a clearer conception of the services and functionality they would like to see being put in place, to enable more productive discussions with funders. As the title of the document indicates, this is a "draft" Strategic Agenda that aims to capture the different stakeholder views about the various proposals. A final version of the Strategic Agenda, with the aim of establishing a HRIC in the context of EHDS and EOSC, is to be delivered at the end of the HealthyCloud project after discussions with stakeholders.

In three of the suggestions (numbers 6-8) the HRIC service is explicitly envisaged as being within EOSC. This may seem to contradict the assertion that we cannot make assumptions now about funding and structures. There are, however, several reasons for leaning towards EOSC as the natural home of at least some HRIC services:

- In initial discussions around HRIC it was usually (and often still is) referred to as "EOSC Health", though to be fair at that stage the EHDS was much less well defined.
- Several of the initiatives within EOSC projects, such as the work on discovery metadata, already overlap with some for the services described for the HRIC.
- The perception that EOSC, with its broad remit within science, would be a more natural and flexible "home" for an interdisciplinary structure like a HRIC.
- The concern, as described in 1.3(g), expressed by some members of HealthyCloud (certainly not all) that the EHDS will inevitably be too focused on developing systems to manage routine health data to have sufficient capacity to handle a wide variety of research generated data.

Having said all of that, if it turns out that EHDS has the funds to develop services and EOSC does not, or if EHDS development proceeds at a faster pace than EOSC, then we see no difficulty at all in embedding services within EHDS, so that rather than a HRIC being "EOSC Health" it becomes "EHDS Research". The requirement is that the identified services are developed and maintained, and the "banner" under which that takes place is not important. The assumption is that the various health-related research communities will also not care, in any significant way, about how those services are funded, "branded", or organised - their concern will be with the usefulness of the service.

The final part of this section provides a brief summary of the proposals, including a one page summary table. To be clear, this draft Strategic Agenda is not proposing a clear "roadmap" for a HRIC and the services listed below are not, at this stage, proposals for funding. Instead, the document is attempting to collect together the various conceptions that have been attached to the idea of a HRIC, to help stakeholders (including of course the HealthyCloud participants themselves) identify, clarify and prioritise the services that should be included within the umbrella of a HRIC.

3.1 A monitoring service for health-related research

One outcome from the HealthyCloud project would be for the deliverables to be collected from the various WPs, including the Strategic Agenda from WP8, and – hopefully – then used to inform future Calls within Horizon Europe, but without the overhead of creating any new entity, or developing any HRIC specific services.

That would be the simplest and cheapest option, but it means that there would then not be a mechanism available to the community to monitor progress towards implementation of services that had been identified as needed or useful. Discussions within HealthyCloud have confirmed that the concerns expressed in the 2018 workshop are still present, and there is no obvious reason that they will decrease in future years – if anything, the concern is that they may increase.

It could therefore be useful – as a minimum – to provide some high-level monitoring functionality through a HRIC, based within the health research communities and providing periodic reports to those communities, and the European Commission, about the "state of play" of health-related research and the progress towards goals that had been previously identified.

The mechanism for such monitoring activity could be as simple as periodic repeats (for example, every 2 or 3 years) of the original workshop with a corresponding report being generated. It could involve, in addition or instead, periodic structured discussions between representatives of funders – especially the European Commission - and the research communities, again with an associated publication. It could involve a standing group within EOSC of health researchers orchestrating surveys and providing periodic reports back to the governance structures of EOSC. Or that working group could be made broader, and include – for instance – representatives of industry as well as a wide range of academic sectors. It could also include horizon scanning using an appropriate discovery tool.

There are clearly a wide range of detailed options possible, and most are relatively cheap. But selecting one would create what might be considered to be a "minimal HRIC" – a mechanism for maintaining the sort of input originally organised in 2018, that has now been expanded by the HealthyCloud project, in future years. The HRIC monitoring service would then have a relatively long life time (though it still might be part of a larger organisation such as EOSC) and an initial core purpose, which could be augmented by some of the services described below.

3.2 A legal / regulatory guidance service

A recurrent theme in recent years has been the difficulty in obtaining clear guidance, or descriptions of best practice, to ensure compliance with legislation and regulations around re-use of personal data, especially as regards the GDPR and other (e.g. national) data protection regulations. As described previously, this situation will change in the next few years but, while the broad "direction of travel" is known, the exact details, especially with regard to the use of health-related data for research, are unclear.

In these circumstances, a possible HRIC service could be the creation of a centre of expertise and resources around legal and regulatory issues, partly to ensure that material to help researchers interpret and comply with regulations were in place, but partly also to ensure that the views of researchers around these issues and the impact on their work were known and fed into the broader debates on the evolving legal and regulatory framework.

Detailed activities could include:

- Developing mechanisms to monitor and analyse the impact of legislative frameworks on the secondary use of data within health-related research, and continuing to identify barriers to data sharing.
- Collecting the views of health researchers about the impact of the legislative framework on their work and their ability to obtain or share data, and ensuring the dissemination of those views.
- Working with expert groups, such as the European Data Protection Board (EDPB), to develop definitions and related guidance to minimise uncertainties, as the legal framework continues to evolve.
- Again working with expert groups and stakeholders, developing international Codes of Conduct that can take into account specific features of different data processing sectors, provide greater guidance to researchers and others, and help to align interpretation and practice across different countries.
- Working with other interested parties, develop model data transfer and data use agreements that can be applied or adapted to a variety of data sharing scenarios, and which are compliant with all EEA legislations.
- Clarifying the legal implications and advantages of analysis in-situ, SPEs, federated analysis, and other innovative models of data access.

How such a service could and should be funded, how it would be "branded", whether it would be a temporary project-based activity or a semi-permanent feature of the landscape, and whether or not it would require a core staff group are funding

decisions for the future. The key question here is whether such a service would be useful, and if so whether it would be a legitimate part of HRIC.

3.3 A training service for researchers

This potential service is similar to that described above, but with the focus on more technical aspects of training rather than regulatory compliance. The expertise and personnel required would therefore be quite different.

There are a variety of potential training "areas" where materials, including possibly courses, could be created, though exact requirements and priorities are liable to vary over time:

- Helping researchers make use of HRIC / EOSC / EHDS (etc.) services and catalogues, to help ensure people were aware of their scope and utility, and could use them effectively.
- Providing support material in relation to preparing data for long term storage and re-use, for example in using de-identification techniques.
- Supporting the use of Secure Processing Environments and / or federated analysis techniques, and the adaptations of analysis that might be necessary.
- Helping researchers understand and use data and metadata standards, including the identification of relevant resources and the practical application of FAIR principles.
- Helping to make cross domain work easier by organising material highlighting the differences in data organisation and metadata within different domains.

Whether or not these and similar training initiatives should or even could be centralised into a single service, or are better handled as extensions of individual projects, or the work of existing research infrastructures, is a question for debate. Again, however, this training activity (one of the five key recommendations from the original 2018 workshop) is a possible HRIC service.

3.4 A metadata standards service

The importance of metadata in supporting the "findability" of data and other resources is widely acknowledged, but so is the problem that we have a large number of different metadata schemas which do not always consider the same objects and attributes. Even within the biomedical sciences there are hundreds of metadata schemas – the situation is made even worse when considering social and environmental sciences as well.

A potential HRIC service in this area could support:

- The development, promotion and application of more consistent discovery metadata schemas, in particular within health-related research. Such schemas should be applied, in particular, to the digital objects available through EOSC.
- The promotion and / or development of suitable discovery metadata schemas for sensitive data. Such schemas need to include reference to levels of risk, (e.g. data de-identification levels) access conditions (e.g. prerequisites demanded of secondary users), and access mechanisms (e.g. who to apply to) where relevant.

- Work, in conjunction with the various other ongoing projects in this area, to enhance linkage of metadata between domains (e.g. health-related research, social sciences, omics).
- Developing mechanisms for the consistent application of PIDs (e.g. DOIs) to digital objects within health-related research. Despite PIDs being routinely available for publications, this is not always the case for other digital objects such as data management plans, protocols, statistical analysis plans, etc.
- The production of software tools and infrastructures that can support all of the above, in particular the easy application of discovery metadata to digital objects, the harvesting of such metadata from different sources, and the production of associated documentation.
- The development of use cases that illustrate how discoverability metadata can be used within catalogues and repository systems to enhance the findability and accessibility of data.

As above the debate here is not whether such services would be useful – they clearly would be and could play a key role in making data more FAIR. Some of these issues are being tackled within other projects (crosswalks between DCAT AP, bioschema.org, DDI, ECRIN metadata schema etc.) and initiatives (e.g., the EOSC Association Taskforce Advisory Group *Metadata and Data Quality*). The question is whether these could and should form a "HRIC service", with a particular focus on health-related research.

3.5 A data interoperability service

Interoperability of data is highly dependent on the use of data standards, and in particular on shared controlled vocabulary systems. The use of data standards in academic clinical research has historically been low, though it has been much higher in industry because of the demands of regulators, especially the FDA. There is a much broader issue here, however, of bridging the gap between the vocabulary systems used in healthcare – which the EHDS proposals might help, in time, to make more consistent – and those used in research, clinical and otherwise.

A HRIC could provide a useful service in this area, by (for example):

- Understanding the levels of use, and types, of data standards in different areas of health-related research.
- Encouraging and supporting the use of the major standard systems within both clinical research (e.g., CDISC) and healthcare (e.g., OMOP, HL7 FHIR). This could include educational materials and input, as well as possible direct support, e.g. by helping researchers apply standards, or by sponsoring research that investigated the costs and benefits of standard use.
- Working with funders and regulators, identifying and helping to apply incentives, resources and pressures that can be useful for promoting interoperability. This includes making use of existing national roadmaps or regional projects intended to increase interoperability.
- Developing infrastructures, systems, resources and tools that can be used to provide *syntactic* interoperability between the systems in use, for example using ETL mechanisms. This should be done in conjunction with the existing initiatives in this area.

- Exploring the best ways to tackle the semantic interoperability problem – including assessing the scale of that problem, examining the applicability, costs, pros and cons of mapping exercises, assessing AI based startups that claim they can provide semantic interoperability, and exploring the feasibility of approaches based on system convergence / aggregation.
- Exploring how licensing models could be made more uniform, and if possible, how standards systems can be made free at the point of use. Making proposals based on this work.
- Maintaining input in global forums around these issues.

The same questions apply as before – is this a legitimate task for a HRIC?, is it already being addressed elsewhere?, what particular contribution could a HRIC make?, etc.

There is also the issue that working to improve interoperability, over and beyond short term mapping exercises, is likely to be a decades-long effort. It may be, for example, that a HRIC could more usefully help to orchestrate and monitor progress in this area rather than be directly involved in the initiatives to improve interoperability.

3.6 An EOSC sensitive data users service

Probably the most common conception of a HRIC has been as "EOSC Health" – a sub-part of EOSC dealing with health-related research and, in particular, sensitive data. But as this and the next two listed proposals attempt to show, what that means in practice is subject to different interpretations.

The most basic incarnation of "EOSC Health" would be as a special interest group within EOSC. Such a group, selected (by some mechanism) to represent researchers working in health-related areas, would try to ensure that developing EOSC systems properly supported the needs of health-related research, and / or could lobby for the inclusion of specific services if they were felt to be missing.

The group would need mechanisms to periodically gather views of researchers about EOSC's support for health-related research – perhaps using formal, structured methods, perhaps more informally. It should also be encouraged to report back on its efforts within EOSC to the research community, and be able to provide a public report addressed directly to EOSC, commenting on the perceived impact of EOSC for health-related research, and identifying areas where challenges remained.

This proposal is similar in some ways to 3.1 above – with the HRIC providing long term monitoring functionality, but in this case it is much more focused on, and embedded within, EOSC. It would not involve the group having any direct responsibility for operations within EOSC.

3.7 An "EOSC Health" catalogue service

As EOSC onboards more resources, and catalogues them for the benefit of users, and as it develops more services, the question arises of whether health-related research needs specialist input, to help identify relevant resources, to support particular types of processing environments, to ensure legal compliance etc.

If it does, then should a HRIC be the part of EOSC that manages, or contributes towards the management of, this part of the EOSC catalogue? Working within the overall policies and structures of EOSC, such a HRIC could, for example:

- Establish mechanisms for identifying and documenting users' functional and computational requirements.
- Identify computational and storage capacities, check availability of resources, set up primary and secondary "data hubs", arrange for the mirroring of resources, and apply monitoring and scheduling systems. Decide on the level of integration between the health-related catalogue and the rest of the EOSC system, perhaps federating the two systems rather than merging them into a single catalogue.
- Develop insight into how federated queries can best be constructed, actioned and stored for possible future use, and identify the implications of such processes for source system APIs.
- Develop mechanisms for "translation" of data into common formats, when it comes from different source systems and uses different file formats and encodings.
- Establishing how user views on system functionality can best be collected and fed back to the HRIC and EOSC.

In this scenario, the HRIC is not a service provider, but it is identifying, and helping to recruit and coordinate, services to EOSC, in the particular domains of health-related research. How this functionality could be integrated within EOSC is not clear – but if there is a need for it should be possible to create the necessary structures, policies and workflows. Again, whether this functionality should be labelled as part of a "HRIC", or simply as "EOSC Health" is open to debate.

3.8 An "EOSC Health" resource service

A HRIC could go beyond curating and managing resources in EOSC, to become a service provider / coordinator in its own right. This could apply in particular to the provision of a network of secure processing environments (SPEs) (also known as Trusted Research Environments or data safe havens), some of which could be pre-existing facilities, some of which might be developed and funded as part of a HRIC infrastructure, and some of which could be part of, or linked to, SPEs made available under the EHDS.

This could involve, amongst other things:

- Identifying existing repositories and SPEs, and identifying possible gaps in the current provision.
- Identifying and investigating the cost models of the best system architectures and infrastructures on which to run HRIC specific services.
- Where necessary, developing and managing additional repositories, for both file based and data platform storage.
- Where necessary, identifying and developing reference implementations for SPEs that can be deployed at data sources for allowing operations with sensitive data "in-situ".
- Developing generally applicable policies and protocols for the storage and use of sensitive data in SPEs.

- Investigating and – where necessary – implementing the role and management of encryption (for data in transit and / or at rest).
- Developing technical protocols for combining analysis of both EHDS stored data and health research data.
- Developing guidance, tools and services, embedding them within the technical infrastructure where appropriate, to help individual researchers or teams protect the privacy of data subjects.
- Implementing and documenting the validation of systems against their identified functional requirements.

Rather than having a possible mosaic of different facilities, with different policies and procedures, managing existing and new SPEs within a single, coordinated structure would allow better management of data, including increased possibilities for federated queries, and should also make it a lot simpler for users.

This particular conception of a HRIC was one of the core ideas expressed in the workshop of 2018, though it is – potentially – also one of the more ambitious and costly versions of a HRIC. It is also not clear at the moment how the introduction of SPEs within EHDS linked infrastructure and elsewhere might affect the use of SPEs for research data – for example whether the same or distinct infrastructures will be required. As before, the key questions are therefore more about the potential value and desirability of such a service, rather than detailed debate about exactly who should fund or manage what.

3.9 A research community interface service, with HealthData@EU

There is no doubt that HealthData@EU, has the potential to create an important and diverse federation of data resources for researchers working in health-related research. But, because the EHDS proposal has not yet been formally adopted, and the system itself will take some years to be fully implemented, there will need to be ongoing debate between researchers and HealthData@EU to maximise its value for research purposes.

For example, data will need to be structured and expressed in ways that are familiar to, and thus most useful to, researchers (or be easily mappable to such forms), application and approval processes for data access will need to be as streamlined as possible, and care will need to be taken that policies, procedures, structures and systems work as well for research derived data as they do for healthcare data. The nature of the de-identification required for both “real world” data used for research purposes, and that generated by research activity, as well as the best place and means of applying such de-identification, is another area where effective liaison with HealthData@EU is required to develop the most efficient solutions and streamline secondary use.

As HealthData@EU develops and publishes increasingly detailed specifications of the services to be established (e.g. the TEHDAS Options for the minimum set of services for secondary use of health data in the EHDS [49]), a dialogue on the interaction between EHDS and health-related research communities becomes more urgent.

The question that then arises is whether a HRIC could usefully be one of the agencies participating in that debate, as the representatives of researchers at a cross disciplinary and transnational level. Exactly how a HRIC would organise this form of representation would need to be decided, but it does offer a mechanism for a communication channel between the EHDS, research infrastructures, and researchers, one of the EHDS's major user groups. While such debate will certainly take place in national fora, and perhaps within particular disciplines, a HRIC based strategy would provide the opportunity to collect and represent views from across the full range of researchers using health data.

3.10 A research community interface service, with the general public

This service was not part of the original set identified in the first HRIC workshop but it has arisen repeatedly in discussions within HealthyCloud. This is the idea that the HRIC acts as part of the interface between researchers and the general public, helping to provide greater understanding of why personal data is important for research, how it is used and the safeguards employed, and helping to build trust around the secondary use of data for research purposes.

Activities in this area could include:

- Working with other organisations, initiatives and projects (e.g., EOSC, EHDS) to develop joint principles of action and ethical frameworks around the re-use of data in health research, and help to ensure that policies and practice were aligned with them.
- Developing mechanisms to promote involvement from patient groups and / or the general public in the work of the HRIC and in data sharing / re-use initiatives generally.
- Providing material and, where appropriate, educational initiatives, to help educate both the press and the general public about the benefits of the research data space and data sharing in general.
- Developing mechanisms for the return of research results to research participants, in a reasonable timeframe, and expressed in terms likely to be understood by most participants.
- Ensuring transparency in secondary use of data, and demonstrating how such use adheres to the relevant legislative and regulatory requirements.

Although not part of the core research function, helping to maintain public trust in research, and in the use of personal data in particular, may be critical not just for the longer-term practice of data sharing but also for the health of populations - for example the COVID-19 pandemic has shown that distrust in science is correlated with low vaccination levels [50]. Unfortunately, only a few "bad episodes" are required to diminish that trust, so it will be important for mechanisms to exist that can help maintain high ethical standards and help explain the use of data in health research more widely. The question is whether a HRIC should be one of those mechanisms.

3.11 Proposals summary

As the summary table below shows, the 10 ideas listed in this document vary in their scope, ambition and likely cost, but all have been raised as potential services within a HRIC. While in some cases they have been proposed, in isolation, as *the* HRIC, in general they have been suggested as one of a cluster of possible components, i.e. as part of a HRIC envisaged as a loose collection of services, all characterised by a direct or indirect usefulness to researchers wanting to more effectively use data in health-related research.

Please note that in the table the “Within EOSC / EHDS” columns are speculative - and that being “possibly within” either does not prevent that service also being in the other infrastructure, or partially outside of either or both.

Some but certainly not all of the possible services directly relate to “Clouds”, in the sense of a computational IT infrastructure. That is not seen as a problem – the criteria for inclusion should be the potential usefulness of the suggestion, not whether it is based in some way on a “cloud”, however defined. In any case, “clouds” of different sorts have now become such a pervasive feature of our personal and professional lives that almost anything can be related to a “cloud”.

Some of the tasks could be addressed, or at least substantially propelled, by a time limited project, but in general the problems that are being tackled are not time limited, so any form of responding service will be required in the longer term. As discussed below this has significant implications for sustainability.

Many of the proposals would also benefit from a small standing staff, if only to carry out administration, organise meetings and collate and edit documents, and assemble and / or coordinate specialist expert groups when required. Once a final cluster of required services is identified then, assuming it includes more than the very basic monitoring function described in proposal 1, the need for a small permanent secretariat and group of core staff will need to be assessed and factored into the costs.

*Table 2: **Summary of Proposals**. This is necessarily very approximate - the time and resources required will depend on the pattern of implementation of any of these services.*

#	Suggested Service	Focus	Time frame	Within EOSC?	Within EHDS?	Overall resourcing
1	A monitoring service for health-related research	monitoring	long	possibly	possibly	low
2	A legal / regulatory guidance service	training & materials	medium / long	possibly	possibly	low / medium
3	A training service for researchers	training & materials	long	possibly	possibly	medium
4	A metadata standards service	data & metadata standards	medium / long	possibly	possibly	medium
5	A data interoperability service	data & metadata standards	long	possibly	possibly	medium
6	An EOSC sensitive data users service	monitoring	long	yes	no	low
7	An “EOSC Health” catalogue service	data services	long	yes	no	high
8	An “EOSC Health” resource service	data services	long	yes	no	high
9	A research community interface service, with EHDS	data services	long	possibly	no	medium
10	A research community interface service, with the general public	training & materials	long	possibly	possibly	medium

4 Sustainability issues

Classically, form follows function. It may be more accurate, however, at least in the context of research infrastructures, to say that form follows funding. Unfortunately, no funding decisions have yet been made about a HRIC or any part of it, so no funds have been allocated. In these circumstances, developing a Strategic Agenda has therefore mostly focused on creating a list of desired services and features, and justifying the

entries in that list, as in sections 1.3 and 3 above. Clarifying and prioritising that list, as has already been made clear, is the first task the stakeholders face.

The second is probably then to investigate the funding opportunities for each entry, recommending the most appropriate, to provide the necessary sustainability for services. One fundamental issue relates to continuity of funding. It is the nature of RIs, (and in some ways the HRIC services look like an embryonic RI, or even collection of RIs) that they require longer term funding to be able to maintain services, and to give users the confidence to invest in using those services. Project funding is good for kicking things off but extended financial support will be necessary for supporting long-term infrastructures such as Secure Processing Environments, or long-term efforts such as improving data interoperability.

With the development of EOSC and EHDS, obtaining funding for a third infrastructure in a similar "space", even a relatively small one such as a HRIC, would probably be very difficult. Aside from funding, there are also the administrative and bureaucratic overheads associated with starting another infrastructure organisation. If HRIC services are to have any form of longevity therefore, most will probably need to be provided as part of another organisation.

As envisaged originally, that organisation, in most cases, was EOSC, and as can be seen from some of the suggestions listed in section 3 that probably remains the most common assumption. We reiterate, however, that we are agnostic about the organisational umbrella or "brand" under which HRIC services are developed - it will depend very much on which organisations have the funding and the motivation to support the services. There is also nothing in theory to prevent some of the services being provided by *both* EOSC and EHDS, with responsibilities divided up in some way, or by other organisations altogether - we seek whatever arrangements best sustain services in the longer term.

Having said that, it is obvious that project funding could also be very useful in some contexts, to tackle immediate needs and demonstrate systems. These could be individual CSA / RIA projects, (for example HORIZON-INFRA-2023-EOSC-01-06: Trusted environments for sensitive data management in EOSC) or a large scale cluster project that could look at several of the issues at the same time, perhaps on the scale of CORBEL or EOSC-Life. In fact, a cross-cluster project might be better – as a HRIC should take a very wide view of the determinants of health and encourage the use of data from the social and environmental sciences (for example) as well as from traditional biomedical sources. As discussed previously, that will require work on the best ways of describing and combining such data, and a cross-cluster project may provide the best opportunity for such work.

As happened with the development of many ESFRI infrastructures, a further preparatory project may be necessary to explore the implementation of different proposals more deeply, for example using working groups (for ECRIN this was done in the ECRIN PPI project that ran from 2008 to 2012). This would allow a detailed plan to be established for the HRIC, taking into consideration the ongoing developments within EOSC and EHDS. The current reflections still appear too premature to generate a detailed agenda, so the Strategic Agenda from HealthyCloud may have to be focused instead on how to plan and develop a HRIC over a longer period. One suggestion is to

plan the HRIC for implementation in 2025. This is probably a more realistic approach and it should lead to a more robust and more useful end product.

A further advantage of this approach is that it would provide more opportunity to investigate costs, perhaps in the context of use cases. Any more specific plan will need to include detailed costing and business planning, so that the costs and benefits of any proposals are more clearly understood.

An interesting model for large scale, longer term initiatives within Horizon Europe is provided by "EU Missions", [51] and it may be that this model could be applicable to a HRIC. There are currently 5 EU missions, dealing with Cancer, Climate Change, Water Quality, Smart Cities and Healthy Soils. In each case, they are intended to [51]:

- be bold, inspirational and widely relevant to society,
- be clearly framed: targeted, measurable and time-bound,
- establish impact-driven but realistic goals,
- mobilise resources on EU, national and local levels,
- link activities across different disciplines and different types of research and innovation,
- make it easier for citizens to understand the value of investments in research and innovation,

all of which have also been identified as necessary attributes of a HRIC.

Within the Cancer mission, for example, there are funding opportunities under the Horizon Europe Programme, the EU4Health Programme, and the Digital Europe Programme, amongst others [52], and a scientific board (a Mission Board) makes recommendations for funding Calls. In a HRIC, the scientific board could include DG Connect, DG RTD and DG Santé (and maybe also IHI) as well as representatives of RIs and national research organisations. A "Health Research and Innovation Mission" could therefore provide much of the high level coordination and monitoring required, while developing and using existing funding channels and infrastructures to deliver the required services.

Whatever the eventual sustainability model, within the current HealthyCloud process it is suggested that the next steps should be:

1. To focus initially on requirements, to define the services seen as required by, and / or beneficial to, data handling in health-related research.
2. To use the provisional list of services presented within this document, refined or extended by stakeholder meetings, as the starting point for those discussions.
3. Once the list of services is defined, to then, and only then, consider in detail how they might be most effectively and efficiently delivered, including costs and financing mechanisms.
4. In particular, to identify which services could be developed within existing infrastructures and projects, and the mechanisms that can be used to ensure that this happens.
5. To then identify the remaining services and functionality, and the associated infrastructure and projects, required to fill the gaps and coordinate the

developments required, continuing to clarify implementation details, costs and funding mechanisms.
6. Implement the HRIC as per the specifications in 4 and 5.

How far the HealthyCloud stakeholder meetings can proceed along these steps, and how many of them would need to be transferred, for example to a later preparatory project as described above, remains to be seen. This document, which is designed to be revised within a succession of stakeholder meetings, can be viewed as the starting point for this process. The final document, due at the end of the project, should aim to describe, as a minimum, the results of the first three tasks listed above. Ideally, some work on tasks 4 and 5 should also be possible, enabling more detailed, concrete proposals to be included in the final Strategic Agenda.

5 References

[1] Aarestrup FM, Albeyatti A, Armitage WJ et al. Towards a European health research and innovation cloud (HRIC). Genome Med 12, 18 (2020). https://doi.org/10.1186/s13073-020-0713-z

[2] Wilkinson M, Dumontier M, Aalbersberg I et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[3] HealthyCloud website: Summary, at https://healthycloud.eu/summary/

[4] Demotes J, Ohmann C, Panagiotopoulou M et al. Discussion Paper for the development of the Strategic Agenda for the Health Research and Innovation Cloud. Version 3, final, 31 January 2022. Available at: https://docs.google.com/document/d/1evuTthBCOT81DCsbQ7qZRZr9Qa5IA1tD/edit?rtpof=true

[5] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A European strategy for data. 19 February 2020

Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066

[6] Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act). 25 November 2020

Available at:

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767

[7] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act). PE/85/2021/REV/1. 30 May 2022

Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868

[8] Data Act: Proposal for a regulation on harmonised rules on fair access to and use of data. 23 February 2022

Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN

[9] Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM/2021/206 final. 21 April 2021

Available at: EUR-Lex - 52021PC0206 - EN - EUR-Lex (europa.eu)

[10] Proposal for a Decision of the European Parliament and of the Council establishing the 2030 Policy Programme “Path to the Digital Decade”. 15 September 2021

Available at:

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0574

[11] Europe's Digital Decade: digital targets for 2030 described at: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_en

[12] Proposal for a regulation of the European Parliament and of the Council on the European Health Data Space. COM(2022) 197/2. 3 May 2022

Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197

[13] Gaia-X Domain Health, Position Paper Version 1.0 2021. Available at:

https://www.bmwk.de/Redaktion/DE/Publikationen/Digitale-Welt/211116-pp-health.pdf?__blob=publicationFile&v=6

[14] Health-X dataLOFT description available at: https://www.health-x.org/home

[15] EOSC News. Available at: https://eosc-portal.eu/media/news

[16] Progress on Open Science: Towards a Shared Research Knowledge System. Final Report of the Open Science Policy Platform. European Commission 2020, available at https://op.europa.eu/en/publication-detail/-/publication/d36f8071-99bd-11ea-aac4-01aa75ed71a1.

[17] The Health Data Hub (France). Available at: https://www.health-data-hub.fr/

[18] Medical Informatics Initiative (Germany). Available at: https://www.medizininformatik-initiative.de/

[19] Finnish social and health data permit authority Findata. Available at: https://findata.fi/en/

[20] Panagiotopoulou M, Yaïche S, Michon A et al., EOSC-Life Public database inventorying the national health databases and registries and describing their access procedures for reuse for research purposes. Zenodo (2022). Available at: https://doi.org/10.5281/zenodo.7148861

[21] Strategy for Digital Health in Spain (published in December 2021). Available at:

https://www.sanidad.gob.es/ciudadanos/pdf/Estrategia_de_Salud_Digital_del_SNS.pdf

[22] EHDEN – Becoming the trusted open science community, built with standardised health data, via a European federated network, at https://www.ehden.eu/.

[23] Data Analysis and Real World Interrogation Network (DARWIN EU), at https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu.

[24] A review commissioned by the Secretary of State for Health and Social Care. Better, broader, safer: Using Health Data for Research and Analysis. April 2022

Available at:
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1067053/goldacre-review-using-health-data-for-research-and-analysis.pdf

[25] Munjal K., Bhatia R. A systematic review of homomorphic encryption and its contributions in healthcare industry. Complex Intell. Syst. (2022). https://doi.org/10.1007/s40747-022-00756-z

[26] Yu Y, Weber G. Balancing Accuracy and Privacy in Federated Queries of Clinical Data Repositories: Algorithm Development and Validation. J Med Internet Res. 2020 Nov 3;22(11):e18735. doi: 10.2196/18735.

[27] Froelicher D, Troncoso-Pastoriza J, Raisaro J et al. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. Nat Commun 12, 5910 (2021). https://doi.org/10.1038/s41467-021-25972-y

[28] Locke S, Bashall A, Al-Adely S, et al. Natural language processing in medicine: A review. Trends in Anaesthesia and Critical Care, Volume 38, 2021, Pages 4-9, https://doi.org/10.1016/j.tacc.2021.02.007.

[29] Zia A, Aziz M, Popa I, Khan SA, Hamedani AF, Asif AR. Artificial Intelligence-Based Medical Data Mining. *Journal of Personalized Medicine*. 2022; 12(9):1359. https://doi.org/10.3390/jpm12091359h.

[30] Wu WT, Li YJ, Feng AZ, Li L, Huang T, Xu AD, Lyu J. Data mining in clinical big data: the frequently used databases, steps, and methodological models. Mil Med Res. 2021 Aug 11;8(1):44. doi: 10.1186/s40779-021-00338-z. PMID: 34380547; PMCID: PMC8356424.

[31] Kinsella CM, Santos PD, Postigo-Hidalgo I, Folgueiras-González A, Passchier TC, Szillat KP, et al. Preparedness needs research: How fundamental science and international collaboration accelerated the response to COVID-19. PLoS Pathog 16(10): e1008902. (2020) https://doi.org/10.1371/journal.ppat.1008902

[32] Besançon L, Peiffer-Smadja N, Segalas C. et al. Open science saves lives: lessons from the COVID-19 pandemic. BMC Med Res Methodol 21, 117 (2021). https://doi.org/10.1186/s12874-021-01304-y

[33] Monkeypox vaccine: WHO urges countries to share data on effectiveness. UN News, 27/07/2022. At https://news.un.org/en/story/2022/07/1123412.

[34] Kamlana A. Your medical records are about to be given away. As GPs, we're fighting back. The Guardian, 03/06/2021, at https://www.theguardian.com/commentisfree/2021/jun/03/gp-nhs-digital-data-patients-records-england.

[35] Pilkington E. Google's secret cache of medical data includes names and full details of millions – whistleblower. The Guardian, 12/11/2019, at https://www.theguardian.com/technology/2019/nov/12/google-medical-data-project-nightingale-secret-transfer-us-health-information.

[36] Coble S. Medical Data of 500,000 French Residents Leaked Online. Info security. 24/02/2021. At https://www.infosecurity-magazine.com/news/500k-french-medical-records-leaked#:~:text=The%20security%20incident%20comes%20after%20an%20announcement%20by,exposed%20in%20this%20attack%20were%20usernames%20and%20passwords.

[37] Questions and answers - EU Health: European Health Data Space (EHDS), at https://ec.europa.eu/commission/presscorner/detail/en/QANDA_22_2712.

[38] National Institutes of Health. Final NIH Policy for Data Management and Sharing, October 2020. Available at: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

[39] Ohmann C, Matei M, Canham S, Panagiotopoulou M, Demotes J, Thomassen G & Blomberg N. EOSC-Life WP14: COVID-19 Repository Data Sharing Policy (Version 1). Zenodo (2021). Available at: https://doi.org/10.5281/zenodo.5519122

[40] Lolfing N. The European Health Data Space Proposal of the European Commission. Bird & Bird. 5 May 2022

Available at: https://www.twobirds.com/en/insights/2022/germany/the-european-health-data-space-proposal-of-the-european-commission

[41] Canham S, Ohmann C, Boiten J-W. et al., EOSC-Life Report on data standards for observational and interventional studies, and interoperability between healthcare and research data. Zenodo (2021). Available at: https://doi.org/10.5281/zenodo.5810612

[42] EOSC-Life website: https://www.eosc-life.eu/

[43] BY-COVID website: https://by-covid.org/

[44] EOSC Future website: https://eoscfuture.eu/

[45] FAIRCORE4EOSC website: https://faircore4eosc.eu/

[46] The European Health Information Portal, website: https://www.healthinformationportal.eu/catalogs/find-data

[47] Welcome to Codex, the Covid-19 Data exchange Platform, website: https://num-codex.de/home

[48] The Maelstrom catalogue. Available at: https://www.maelstrom-research.org/page/catalogue

[49] TEHDAS JA Deliverable 7.1: Options for the minimum set of services for secondary use of health data in the EHDS. Available at: https://tehdas.eu/app/uploads/2022/04/tehdas-options-for-the-minimum-set-of-services-for-secondary-use-of-health-data-in-the-ehds.pdf

[50] Algan Y, Cohen D, Davoine E et al. Trust in scientists in times of pandemic: Panel evidence from 12 countries. Proceedings of the National Academy of Sciences, vol 118, no 40, October 5, 2021. Available at: https://www.pnas.org/doi/10.1073/pnas.2108576118

[51] EU Missions in Horizon Europe. Available at: https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/eu-missions-horizon-europe_en

[52] EU Mission: Cancer. Available at: https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/eu-missions-horizon-europe/cancer_en

6 Acronyms and Abbreviations

AAI	Authentication and Authorization Infrastructure
AI	Artificial Intelligence
API	Application Programming Interface
B1MG	Beyond 1 Million Genomes (*project acronym*)
BY-COVID	Beyond COVID (project acronym)
CDISC	Clinical Data Interchange Standards Consortium
CORBEL	Coordinated Research Infrastructures Building Enduring Life-science services (*project acronym*)
COVID	Coronavirus disease
CSA	Coordination & Support Action
DARWIN	Data Analysis and Real World Interrogation Network (*project acronym*)
DCAT AP	Data Catalog Vocabulary Application Profile
DDI	Data Documentation Initiative
DG RTD	Directorate-General for Research and Innovation
DOI	Digital Object Identifier
EC	European Commission
ECRIN	European Clinical Research Infrastructure Network
ECRIN PPI	European Clinical Research Infrastructure Network - Preparatory Phase for the Infrastructure (*project acronym*)

EDPB	European Data Protection Board
EEA	European Economic Area
EHDEN	European Health Data Evidence Network (*project acronym*)
EHDS	European Health Data Space
EHDS2	European Health Data Space for the secondary use of health data
EHR	Electronic Health Record
EMA	European Medicines Agency
EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures
ETL	Extract, Transform, Load
EU	European Union
FAIR	Findability, Accessibility, Interoperability, and Reusability
FHIR	Fast Healthcare Interoperability Resources
FDA	Food and Drug Administration (U.S.A.)
GDPR	General Data Protection Regulation
HL7	Health Level Seven International
HRIC	Health Research and Innovation Cloud
IHI	Innovative Health Initiative
IT	Information Technology
NIH	National Institutes of Health (U.S.A)

OMOP	Observational Medical Outcomes Partnership
PID	Persistent Identifier
PHIRI	Population Health Information Research Infrastruture (*project acronym*)
RI	Research Infrastructure
RIA	Research and Innovation Action
RWD	Real World Data
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
SPE	Secure Processing Environment
TEHDAS JA	Towards the European Health Data Space Joint Action (*project acronym*)
TRE	Trusted Research Environment
U.S.A.	United States of America
WP	Work Package

7 Appendix

7.1 The European Open Science Cloud (EOSC)

The European Open Science Cloud (EOSC) was envisaged as a federated, globally accessible environment where researchers, innovators, companies and citizens can publish, find and re-use each other's data and tools for research, innovation and educational purposes. This environment should operate under well-defined conditions to ensure trust and safeguard the public interest.

The EOSC aims to accelerate the transition to more effective Open Science and Open Innovation in a Digital Single Market by removing the technical, legislative and human barriers to the re-use of research data and tools, and by supporting access to services, systems and the flow of data across disciplinary, social and geographical borders. The term “European Open Science Cloud” requires some reflection to dispel incorrect associations and clarify boundaries; in fact the term “cloud” is used as a metaphor to help convey the idea of seamlessness and a commons.

European: research and innovation are global. The EOSC cannot be built exclusively in and for Europe. Serious efforts are ongoing to ensure coordinated action with other regions. Europe, being inherently federated, is in a strong position to lead this initiative.

Open: the use of Open in relation to research has been widely discussed over recent years, and it is acknowledged that not all data and tools can be open. There are exceptions to openness, such as confidentially and privacy. Open is also often confused with "for free". Free data and services do not exist. These nuances need to be respected and intelligently open is what is meant, often referring more to accessibility under proper and well defined conditions for all elements of the EOSC.

Science: the use of the term science explicitly includes the arts and humanities, and in fact no current or future discipline should be excluded from the EOSC. In addition the Science Cloud infrastructure should support not only innovative scientific research but also societal innovation and productivity, which takes place predominantly in collaboration between research institutes and the private sector. The EOSC should also support broad societal participation in Open Innovation and Open Science.

Cloud: the term cloud can cause considerable confusion as it has many connotations. It can be misinterpreted to indicate that the EOSC is mostly about hard ICT infrastructure and much less about a commons of data, software, standards, expertise and policy related to data-driven science and innovation.

The EOSC Objectives Tree (Figure 2) depicts the 3 main EOSC objectives and identifies the main problems, barriers and benefits.

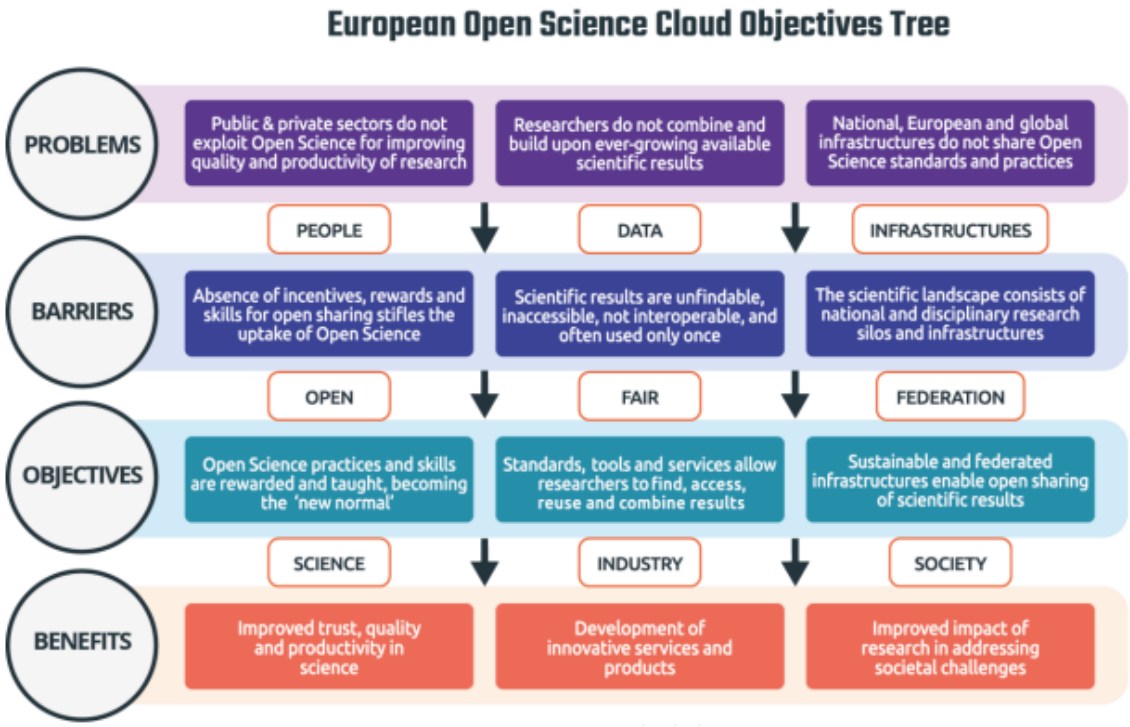

Figure 2: European Open Science Cloud Objectives Tree (Adopted from the EOSC SRIA)

In addition, a set of guiding principles for EOSC has been agreed:

- **Research-community centred**: EOSC will place research at the centre of the initiative and will thus prioritise engagement with research communities to understand their requirements, helping them and ensuring academic sovereignty of research data.
- **Multi-stakeholderism:** EOSC will succeed if and only if it follows a multi-stakeholder approach;
- **Openness:** EOSC will ensure that research artefacts are "as open as possible, as closed as necessary";
- **FAIR principles:** EOSC will assemble research artefacts that are findable, accessible, interoperable and reusable;
- **Federation of infrastructures:** EOSC will federate existing and upcoming research infrastructures;
- **Machine-actionable:** EOSC will strike the right balance between machines and people in delivering the services that will serve the needs of European scientists.

Some immediate implementation challenges that EOSC is facing:

Identifiers: The persistence of the identity of digital objects and stability of references to those objects are essential to sustaining a trusted distributed research ecosystem that supports verifiable and reusable research.

Metadata and ontologies: An overarching, coordinated approach is required, forming the basis for interoperability.

FAIR metrics and certification: Existing work on FAIR metrics and certification should be extended to ensure applicability across disciplines and support their implementation. FAIR assessments must be inclusive and progressive, and take the specific research context and needs into account.

Authentication and authorisation infrastructure (AAI): The purpose of AAI in EOSC is to allow identified scientists to (re)use identifiable documents, data and software, and exploit identified services, while enabling high-trust collaborations to be established and maintained with little or no friction to the end user. The goal is to build a foundation for AAI that will ensure long-term availability of the aspects of digital identity that are unique to scientific collaborations.

User environments: Throughout the distributed, federated and clustered architecture of the EOSC ecosystem, the user environments must meet the users' requirements and expectations, particularly with regard to discovery and composability of resources.

Resource provider environments: As a federation built out of many independent organisations and resource providers – a system of systems – EOSC should be inclusive rather than selective. The added value of EOSC exists only when as many as possible of the resource providers serving the scientific community can enter and offer resources.

EOSC Interoperability Framework: Achieving a good level of technical, semantic, organisational and legal interoperability within EOSC is essential to federate services and provide added value for users, across disciplines, countries and sectors.

Within EOSC there are **seven action areas** relating to the social, financial, legal, educational, cultural challenges and prerequisites referred to as *the boundary conditions*:

1) **Rules of Participation**: A process of change in the research environment is required to adopt Open Science practices, make digital research objects FAIR and federate research data infrastructures. The Rules of Participation (RoP) provide transparent and consistent terms for involvement in EOSC, helping to build the trust and confidence required to support and deliver this process of change.
2) **Landscape monitoring:** Sustainable long-term monitoring of EOSC landscape developments – the infrastructures, initiatives, investments and policies at national and institutional levels – is required to to allow informed decisions on EOSC.
3) **Funding models:** Viable funding models are an essential element of ensuring an operational, scalable and sustainable EOSC ecosystem.
4) **Skills and training:** In order to leverage the potential of EOSC for open and data-intensive research, a key challenge for Europe is to ensure the availability of highly and appropriately skilled people. The vision of a strong EOSC ecosystem that exploits digital technologies and has data and software at its core necessitates a comprehensive skills and education strategy.
5) **Rewards and recognition:** A culture change needs to be realised in the way scientists are appraised, by looking at their broader contribution to education, research, impact and leadership. A responsible rewards and recognition system is a catalyst to foster good research practice and quality in terms of content, openness, scientific integrity and contribution to society.
6) **Communication:** EOSC's diversity of stakeholders requires a communication policy that meets the different needs of each group, providing clarity on the why, how and what of EOSC, and sending out its messages in a consistent way.
7) **Widening to public and private sectors and going global:** To successfully extend the EOSC ecosystem beyond the core research community, EOSC must demonstrate value and impact that is relevant and meaningful to the diverse groups belonging to the broader public sector and to the private sector. In parallel, there is clearly a global dimension to EOSC, a common vision that enables Europe to enhance scientific collaboration with other parts of the world and drive a cultural change towards Open Science, which EOSC must respect and exploit to maximise its potential impacts.

To advance on these aspects EOSC has created six working groups (WGs) consisting of experts from the EOSC projects and stakeholder community: WG Architecture is defining a technical framework to enable and sustain an evolving EOSC federation of systems, including application programming interfaces (APIs), authentication and authorisation infrastructure (AAI), and persistent identifiers (PIDs). WG FAIR is defining requirements for developing, assessing and certifying EOSC services in order to foster cross-disciplinary interoperability through FAIR. WG Landscape is mapping the landscape and readiness of existing research infrastructures in Europe that could be connected to EOSC. WG Rules of Participation is designing the rules to define the rights and obligations governing transactions between EOSC users, providers and operators. WG Skills & Training is providing a framework for a sustainable training infrastructure to support the uptake of EOSC. Finally, WG Sustainability is providing

recommendations on the implementation of a scalable and sustainable EOSC, including business models, integration of national infrastructures, and legal models for EOSC. Different thematic Task Forces (TF) complement the work of the WGs, e.g. TF FAIR Metrics and Data Quality, TF Semantic Interoperability etc.

The implementation of the EOSC is based on a long-term process of alignment and coordination pursued by the European Commission since 2015 with the many and diverse stakeholders of the European research landscape. In the initial phase of implementation (2018-2020), the European Commission invested around €250 million to prototype components of the EOSC through Calls for projects under Horizon 2020. The current phase of implementation (2021-2030), is taking place in the context of the EOSC European co-programmed partnership launched at the Research and Innovation Days 2021 and according to a Strategic Research and Innovation Agenda (SRIA) which is co-developed with the entire EOSC community. EOSC is transitioning to a more stakeholder-driven approach with a shared vision, common objectives and complementary contributions at European, national and institutional levels. A co-investment (with in kind and financial contributions) by the EU and non-EU partners of at least €1 billion is foreseen for the next 7 years.

The EOSC is recognised by the Council of the European Union among the 20 actions of the policy agenda 2022-2024 of the European Research Area (ERA) with the specific objective to deepen open science practices in Europe. It is also recognised as the "science, research and innovation data space" which will be fully articulated with the other sectoral data spaces defined in the European strategy for data.

Overall progress is steered by a new EOSC tripartite governance involving the EU represented by the European Commission, the participating countries represented in the EOSC Steering Board and the research community represented by the EOSC Association.

Key references

- EOSC portal: https://eosc-portal.eu/
- EOSC Association website: https://eosc.eu/
- EOSC Strategic Research and Innovation Agenda version 1.0 (SRIA):

  https://op.europa.eu/en/searchresults?p_p_id=eu_europa_publications_portlet_search_executor_SearchExecutorPortlet_INSTANCE_q8EzsBteHybf&p_p_lifecycle=1&p_p_state=normal&queryText=EOSC+Strategic+Research+and+Innovation+Agenda+version+1.0&facet.collection=EUPub&startRow=1&resultsPerPage=10&SEARCH_TYPE=SIMPLE

- Digital skills for FAIR and Open Science. Report from the EOSC Executive Board Skills and Training Working Group:

  https://op.europa.eu/en/publication-detail/-/publication/af7f7807-6ce1-11eb-aeb5-01aa75ed71a1/language-en/format-PDF/source-190694287

- EOSC architecture working group view on the minimum viable EOSC. Report from the EOSC Executive Board Working Group (WG) Architecture:

  https://op.europa.eu/en/publication-detail/-/publication/91fc0324-6b50-11ebaeb5-01aa75ed71a1/language-en/format-PDF/source-190574886

- EOSC interoperability framework. Report from the EOSC Executive Board Working Groups FAIR and Architecture:

  https://op.europa.eu/en/publication-detail/-/publication/d787ea54-6a87-11ebaeb5-01aa75ed71a1/language-en/format-PDF/source-190578229

- EOSC Authentication and Authorization Infrastructure (AAI). Report from the EOSC Executive Board (EB) Working Group (WG) Architecture AAI Task Force (TF):

  https://op.europa.eu/en/publication-detail/-/publication/d1bc3702-61e5-11ebaeb5-01aa75ed71a1/language-en/format-PDF/source-189583372

- Rules of Participation (RoP):

  https://op.europa.eu/en/publication-detail/-/publication/a96d6233-554e-11ebb59f-01aa75ed71a1/language-en/format-PDF/source-189583337

- Recommendations on FAIR metrics for EOSC:

  https://op.europa.eu/en/publication-detail/-/publication/ced147c9-53c0-11eb-b59f01aa75ed71a1/language-en/format-PDF/source-184008165

- Recommendations on certifying services required to enable FAIR within EOSC:

  https://op.europa.eu/en/publication-detail/-/publication/70aa74b5-53bf-11eb-b59f01aa75ed71a1/language-en/format-PDF/source-184009543

- Persistent Identifiers (PID) architecture for the EOSC. Report from the EOSC EB WG Architecture PID Task Force (TF):

  https://op.europa.eu/en/publication-detail/-/publication/3136c3e6-4f07-11eb-b59f01aa75ed71a1/language-en/format-PDF/source-184010810

7.2 The European Health Data Space (EHDS)

As part of the European Strategy for data, and in order to unleash the full potential of health data, the European Commission published in May 2022 a proposal for the European Health Data Space (EHDS) regulation. This proposal builds further on the General Data Protection Regulation (GDPR), the proposed Data Governance Act, draft Data Act and the Network and Information Systems Directive.

The European Health Data Space is envisaged as a health specific ecosystem comprised of rules, common standards and practices, infrastructures and a governance framework that aims at:

- Empowering individuals through increased digital access to and control of their electronic personal health data, at national level and EU-wide, and support to their free movement, as well as fostering a genuine single market for electronic health record systems, relevant medical devices and high risk AI systems (primary use of data).
- Providing a consistent, trustworthy and efficient set-up for the use of health data for research, innovation, policy-making and regulatory activities (secondary use of data).

Secondary use is described in Chapter IV of the proposed regulation and the following categories of data are concerned:

a) EHRs;
b) data impacting on health, including social, environmental behavioural determinants of health;
c) relevant pathogen genomic data, impacting on human health;
d) health-related administrative data, including claims and reimbursement data;
e) human genetic, genomic and proteomic data;
f) person generated electronic health data, including medical devices, wellness applications or other digital health applications;
g) identification data related to health professionals involved in the treatment of a natural person;
h) population wide health data registries (public health registries);
i) electronic health data from medical registries for specific diseases;
j) electronic health data from clinical trials;
k) electronic health data from medical devices and from registries for medicinal products and medical devices;
l) research cohorts, questionnaires and surveys related to health;
m) electronic health data from biobanks and dedicated databases;
n) electronic data related to insurance status, professional status, education, lifestyle, wellness and behaviour data relevant to health;
o) electronic health data containing various improvements such as correction, annotation, enrichment received by the data holder following a processing based on a data permit.

The regulation sets out permitted purposes for secondary use of health data. Whether the data was initially collected for primary use or for secondary use is irrelevant in this regard. The permitted purposes of processing include among other things, activities for

reasons of public interest in the area of public and occupational health (e.g. protection against serious cross-border health threats), support of health-related research and innovation activities, training, testing and evaluating algorithms and education and teaching activities.

The regulation also prohibits certain types of processing such as taking decisions detrimental to a natural person or excluding them/modifying their insurance contracts and premiums, advertising or marketing activities towards health professionals and organisations or natural persons; developing products or services that may harm individuals and societies at large (e.g. illicit drugs).

Any secondary use of health data requires prior approval by a competent body. Member States need to designate one or more Health Data Access Bodies to ensure the data is made available to data users after the request has been granted and to maintain an administrative system to record and process data access requests, data inquiries and data sharing approvals.

The EHDS proposes a number of different ways to ensure and demonstrate health data quality and utility for secondary use. These include establishing a European Union "data quality and utility" label, using metadata catalogues and source information, harmonised technical and data management processes, and transparency around access, provision, and data enrichment. The different national datasets will be interconnected and linked across the EU by the Commission through a publicly available "EU Datasets Catalogue". The specifics though (e.g. setting out the minimum information elements that data holders are to provide for datasets and their characteristics; determining the minimum specifications for cross-border datasets for secondary use of electronic health data; setting out the visual characteristics and technical specifications of the "data quality and utility label") remain to be clarified by the Commission by means of implementing acts.

Given the sensitivity of health data, the Health Data Access Bodies will be providing access to anonymised data. Where the purpose of the data user's processing cannot be achieved with anonymised data, the Health Data Access Bodies will consider providing access to electronic health data in pseudonymised format but the information necessary to reverse the pseudonymisation will remain available only to the health data access body. Re-identification attempts by data users are prohibited by law.

To improve data interoperability, the draft Regulation places special requirements on EHR systems, (systems used in connection with electronic health records which are intended by their manufacturer for the primary use of prioritised electronic health data). In particular, EHR systems may only be placed on the market and put into operation if the specific requirements of the EHDS are met. These are primarily taken from the criteria listed in Annex II Section 2, which the Commission intends to specify further by means of implementing acts. Currently, the interoperability requirements remain quite "vague" and "broad":

For example (as stated in Annex II, Section 2 of the Regulation):

- *"An EHR system shall allow personal electronic health data to be shared between health professionals or other entities from the health system, and between health professionals and patient or health professional portals in a commonly used electronic interoperable format, which includes, inter-alia, dataset content, data structures, formats, vocabularies, taxonomies, exchange formats, standards, specifications, profiles for exchange and code lists, thus enabling system to system communication."*
- *"An EHR system that includes a functionality for entering structured personal electronic health data shall enable the entry of data structured in a structured way that supports the data sharing in a structured, commonly used and machine-readable format, enabling system-to-system communication."*
- *"An EHR system shall not include features that prohibit, restrict or place undue burden on authorised access, personal electronic health data sharing, or use of personal electronic health data for permitted purposes."*

An important goal of the EHDS is the creation of a powerful infrastructure to facilitate cross-border healthcare and secondary use by interconnecting the authorised participants. Particular innovations in this aspect include the establishment of the EHDS Board, a cross-border infrastructure for the primary use of electronic health data (MyHealth@EU) and a cross-border infrastructure for the secondary use of electronic health data (HealthData@EU).

Authorised participants in HealthData@EU could be Health Data Access Bodies, research infrastructures established as an European Research Infrastructure Consortium ("ERIC"), as well as other types of entities, including infrastructures under the European Strategy Forum on Research Infrastructures (ESFRI), infrastructures federated under the European Open Science Cloud (EOSC). Other authorised participants should obtain the approval of the joint controllership group for joining HealthData@EU.

The European Parliament and the Council are currently discussing the draft legislation.

One of the European projects providing input to the EHDS development is the TEHDAS Joint Action. The project started on 1 February 2021 and will run until 1 August 2023.

The claimed objectives of the project are:

- engaging other European projects and policymakers in a dialogue about the European Health Data Space;
- ensuring sustainability of the secondary use of health data in Europe;
- developing a governance model for cross-border co-operation in the secondary use of health data between European countries;
- promoting the reliability and compatibility of and access to health data for secondary use;
- clarifying the role of individuals in the secondary use of health data and including them in dialogue about the use of health data for research and policymaking.

Figure 3: Work packages of the Joint Action TEHDAS (taken from and more info at: https://tehdas.eu/packages/)

The TEHDAS project has produced a series of documents aiming to inform the EHDS proposal; among them *an overview of data altruism structures and functions for the EHDS*, *data quality recommendations*, *recommendations on data interoperability*, *minimum technical services for the EHDS* and so on. All the current project results are available here: https://tehdas.eu/results/.

More recently, the EHDS2 pilot project led by the French Health Data Hub was launched with the objective to address the challenges surrounding access to health data throughout the EU and to open new perspectives to research and innovation feeding into the EHDS proposal. Currently, several research use cases involving ELIXIR, BBMRI, PHIRI, EMA-DARWIN, and the ECDC are being set up.