AIDAVA
Horizon Europe #101057062

MyData

To solve the problem of **interoperability** and **quality** of health data, we need a structured **roadmap**, involving **individuals** as active stakeholders.

August 2024

Dr Isabelle de Zegher, MD, MSc,
Senior Advisor Health MyData,
Member of the Board of Directors SNOMED International,
Clinical coordinator AIDAVA.

Prof Remzi Celebi, PhD,
Assistant Professor Maastricht University Department of Advanced Computing Sciences,
Technical coordinator AIDAVA.

Table of Contents

TAGS: Health, FAIR, data quality, data interoperability, personal health knowledge graph

Executive Summary

Interoperability, specifically data interoperability, in healthcare, remains a critical yet unresolved issue despite numerous attempts over the past 25 years. We need to learn from the past and develop an alternative strategy with a **structured roadmap** to address this issue, in the same way that we tackled complex problems such as cancer.

It is widely acknowledged that health data sources are heterogeneous in terms of structure and standards, with varying data quality and often limited or poor documentation (metadata). Whilst (too) many standards exist and multiple initiatives aim to ensure interoperability and reusability in secondary use of health data, there is currently no solution that guarantees that the health record of a single individual - the main source for secondary datasets - is consistent, correct, interoperable and reusable. We believe that **data interoperability can be solved** *if* ***each*** *citizen has a personal, high-quality, integrated health record compliant with a* ***global data sharing standard*** *- encompassing relevant existing data standards - that supports seamless reuse.* Collective data interoperability can be achieved by ensuring that each patient's data is compliant with the standard data sharing format in an increasingly digital healthcare ecosystem. This will not only benefit patients and their care providers but also help reduce biases and mistakes in the AI tools using these health data.

This is an ambitious objective. Research from the Horizon Europe project AIDAVA, combined with MyData experience on data intermediaries, suggests that this could be achieved - at a sustainable cost - through three phases, with the support of policy makers.

1. Establish foundational components of interoperability including 1) documentation of each data source, 2) bringing together existing standards under one ontological roof (i.e., upper level healthcare ontology) and 3) continuous improvement of certified AI tools - such as entity linking and multilingual Large Language Model (LLM) - that can automatically transform heterogenous data into the common semantic standard.
2. Develop process and AI based tools to map data sources with the ontology standard and enable automatic transformation of these data sources.
3. Provide individuals with intelligent digital assistants (like the AIDAVA prototype) to seamlessly derive their interoperable and reusable health record, for the benefits of the patients, providers and the research community.

Initial results from AIDAVA, including evaluation of the prototype with patients over 4 clinical sites, demonstrate that this is feasible. To make AIDAVA-like solutions acceptable in real life settings and sustainable, we need to ensure that the aforementioned foundational components are in place.

The issue

Interoperability has been unsolved for more than 25 years.
It is time we follow a different - individuals' centric - approach in managing health care data, and develop a structured roadmap to solve the issue. The roadmap should include both data interoperability and data quality aspects.

There is wide recognition that interoperable, quality data is a critical component in solving many issues faced by current health care systems: lack of coordination and integrated care, increased providers' burnout, poor patient access to services and experience, risks to patient safety, inefficient & costly clinical care with redundant procedures, limited secondary decision making in providing value-based care, difficult and costly access to data for research and policy-making... Yet our data are spread across systems, heterogenous and of questionable quality[1]. Making health data FAIR[2] should therefore be a core objective of all healthcare systems. This document focuses on the "I" in FAIR: interoperability, and more specifically data interoperability.

Interoperability is '*the ability of two or more systems or components to exchange information and to use the information that has been exchanged*'[3]. The challenge of interoperability is not unique to healthcare, and common models for interoperability are proposed across academic literature in healthcare as in other domains. The European Commission presented their viewpoints on an interoperability framework[4] to facilitate digital transformation for citizen empowerment across sectors. The Healthcare Information and Management Systems Society (HIMSS), a society committed to reforming the global health ecosystem through information technology, defines four levels of interoperability[5]. The Network of the National Library of Medicine (NLM) defines the more specific issue of ***data interoperability***[6] as the ways in which data is formatted, that allow diverse datasets to be integrated or aggregated in ***meaningful ways.***

While high level frameworks help to classify the issues, the data interoperability problem has not been solved for more than 25 years in healthcare - and other sectors - despite interoperability roadmaps active in different countries. The complexity of the issue increases with data reuse shifting from 'point-to-point', well defined data exchanges, to 'many-to-many' just in time data sharing, consented by individuals. We argue that it is time to work on a different approach working with patients and citizens as critical stakeholders of

[1] As a patient, the main reason to be in charge of my health data is to ensure its QUALITY (https://mydata.org/2024/05/27)
[2] https://www.nature.com/articles/sdata201618
[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6153178/
[4] https://ec.europa.eu/isa2/sites/isa/files/eif_brochure_final.pdf
[5] https://www.himss.org/resources/interoperability-healthcare
[6] https://www.nnlm.gov/guides/data-glossary/data-interoperability

an individual' centred roadmap, building on lessons learned from previous initiatives and on the emergence of new technologies such as neurosymbolic AI[7].

This document focuses on data interoperability. It does not consider the other pillars needed to achieve full interoperability and addressed at length by authorities, in EU and outside EU, such as

- technical aspects of data transfer (for point-to-point data exchanges or for data sharing), including API specifications, communication protocols, and cybersecurity;
- supportive business and regulatory functions including funding, skills and education, consensus building,..;
- governance to ensure long term sustainability.

Health data management today

Health data typically come from various sources, including hospitals (accounting for approximately 40% of health data), GP practices, home and social care, clinical trials conducted by pharmaceutical and medical devices companies, surveys from public health authorities and increasingly data collected by patients through devices and patient-reported outcomes applications. **All these data are collected for different purposes using different standards** related to their primary use. For instance, care data collection focuses on a single encounter at a specific point in time for an individual patient ('vertical data'). In contrast, clinical research requires collection of specific parameters over time for as many subjects as possible ('horizontal data').
Since the late 80's, multiple data standards have been developed for clinical care and clinical research. A few of these including HL7 (v2, V3, CDA, FHIR), SNOMED, LOINC, ISO 13606, openEHR, CDISC, MEDDRA … are still in use as they rely on strong semantic foundations, often in the form of ontologies.

Most health information systems in use in hospitals are outdated, using legacy designs and built on old technologies (Epic – 1979; Cerner – 1976; McKesson – 1960; MEDITECH – 1969; Allscripts – 1986; SAP IS-H – early 90s and being retired in 2030; many proprietary systems from the 90s). The underlying data models are rarely documented and not easily accessible; extraction of data for reuse is a recurrent and expensive struggle, independently of data privacy considerations. Upgrade of these legacy systems is considered as the top challenge by health care executives in a 2023 survey by McKinsey[8].

[7] https://arxiv.org/abs/2305.00813
[8] https://www.mckinsey.com/industries/healthcare/our-insights/digital-transformation-health-systems-investment-priorities

Data sources within healthcare organisations are of suboptimal quality, though there are no formal metrics publicly reported of large scale quality assessments. One study[9] identified more than 40% redundancies in clinical notes, mainly resulting from copy paste behaviour of junior staff. Another study[10] demonstrated that up to 10% of health records would have errors that can negatively impact the patient and health decision-making. As source data can be represented in various formats, tools for addressing these quality issues and enhancing quality are not easily scalable to handle the diverse, distributed and increasing volume of data, and are rarely in use.

Ongoing initiatives toward interoperability

Authorities across regions have developed health interoperability roadmaps, including legal, organisational, technical and semantic aspects.

Most specifically the US ONC-HIT initiated their roadmap in 2015 with a a 10-year vision that remains fully applicable in 2024: *'a learning health system where individuals are at the centre of their care; where providers have a seamless ability to securely access and use health information from different sources; where an individual's health information is not limited to what is stored in electronic health records (EHRs), but includes information from many different sources (including technologies that individuals use) and portrays a longitudinal picture of their health, not just episodes of care; where diagnostic tests are only repeated when necessary, because the information is readily available; and where public health agencies and researchers can rapidly learn, develop, and deliver cutting edge treatments*.'

While the vision of learning health systems and the proposed approach are similar across countries, the related roadmaps fall short at different levels. In the table below, we compared the roadmaps of 4 regions around critical aspects of data interoperability.

- The European Health Data Space (EHDS) regulation in EU[11],
- The Health Data, Technology, and Interoperability: Patient Engagement, Information Sharing, and Public Health Interoperability (HTI-2). Propose rule in the US[12]
- The shared Pan-Canadian Interoperability Roadmap in Canada[13] and
- The National Healthcare Interoperability Plan in Australia[14]

[9] https://www.sciencedirect.com/science/article/pii/S1532046421002677#s0115
[10] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9983735/
[11] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0197
[12] https://www.healthit.gov/topic/laws-regulation-and-policy/health-data-technology-and-interoperability-certification-program
[13] https://www.infoway-inforoute.ca/en/component/edocman/6444-connecting-you-to-modern-health-care-shared-pan-canadian-interoperability-roadmap/view-document?Itemid=103
[14] https://www.digitalhealth.gov.au/about-us/strategies-and-plans/national-healthcare-interoperability-plan

	EU	US	Canada	Australia
Digital Identity	eIDAS/ DIgital Wallet	?	Digital identity	Healthcare Identifiers Service
Patient engagement	passive access to national health portals	passive access	passive access (extend longitudinal record)	passive - access to MyHealth Record
Interoperable Data Elements	Emerging xShare core data set (extension of IPS data elements)	USCDI[15]	Pan-Canadian Health Data Content Framework - p-CHDCF (aligned with FHIR)[16]	?
Terminology/Data Standards	HL7, SNOMED, LOINC	HL7, SNOMED, LOINC	HL7, SNOMED, LOINC, I	HL7, SNOMED, LOINC, (AMT,METEOR)
Providers directory	Proposed Registry of healthcare providers	?	?	Health service directories
Data Exchange/ API	FHIR EEHRxF (start with 6 standards)	multiple certified FHIR profiles	FHIR IPS	ePrescribing
Catalogue of dataset for secondary use	EHDS Dataset catalogue (+ article 33 - minimum dataset)	?	?	?
Data Sharing approach	Data Altruism/ data donation	? (on demand data exchange)	? (on demand data exchange)	In development (API Gateway information exchange, consent management,..)

Table 1. Comparison of data interoperability component in the interoperability roadmap

Typically authorities made major progress in agreeing on standards and on point-to-point data exchanges; there has been as well major efforts (not documented in the table) on governance around health data management. There are however major weaknesses across all roadmaps.

1. **Patients (and citizens) "engagement" is passive**. They can access their healthcare data, which are spread across different organisations and systems, but there is a lack of tools to integrate their personal data - from personal apps and medical devices - and to increase the quality of this data (missing data, errors, redundancies with inconsistencies…). Current approaches overlook the potential of patients and citizens as contributors and curators to improve the quality of their own health and the health of their loved ones. By not engaging their knowledge and motivation, we miss a massive workforce to improve health data[17]. The difficulty, however, is that patients and citizens need appropriate tools at their level of health and digital literacy to be effective.
2. **Current focus on interoperability is on population data,** without taking into account the critical need for each patient to have a complete, correct, longitudinal health record. As

[15] https://www.healthit.gov/isp/united-states-core-data-interoperability-uscdi#uscdi-v4

[16] https://www.cihi.ca/sites/default/files/document/pan-canadian-health-data-content-framework-data-content-standard-en.pdf

[17] A internet survey made with 250 citizens during the AIDAVA project Deliverable 1.2, showed that 56% of them were ready to increase the quality of their dossier with a tool like AIDAVA; this number increases to 81% when they have an active medical dossier

an example, in the Canadian Interoperability Landscape Study[18] 82% of clinicians mentioned that they do not always have a summary of the care of their patients outside of their practice setting.

3. None of the roadmaps **address data quality**, there are no metrics provided, while several studies mentioned above confirm the suboptimal quality of personal health records. Errors can be smoothed at population level, yet there is always a risk of bias and errors.
4. The main efforts focus on **defining a list of common data elements (CDE)** that is expected to be enforced across stakeholders. While this 'divide-and-conquer' approach offers some benefits, it is ultimately unsustainable as it cannot replace a complete longitudinal health record or foresee the future needs of research. In addition, the CDE list needs to be constantly maintained/updated through consensus, which can be challenging, and each new version has to be rolled out, resulting in additional costs for healthcare organisations, which are already struggling financially.
5. None of the proposed approaches will allow **optimising current practices of recurrent mapping of data sources into a target format for secondary data use.** This involves linking heterogeneous concepts defined in the source data, with concepts specified in the target format. If different human beings execute this task, there is a potential for different interpretations of source data and therefore risk of non-interoperability; also the same data sources can be potentially mapped several times for different data uses. In addition, this is a workload intensive, mostly manual, task which typically takes ⅔ of the work, while ⅓ is then dedicated to the analysis itself. As a result, the generation of high-quality datasets for secondary use is currently a challenge and is expected to remain so in the future.

How to solve health data interoperability

What is the end goal

Our vision is similar to the one proposed in the learning health system 2015 by US ONC-HIT, with a stronger focus on the individuals 'at the centre' and in control of their data as advocated in the original MyData white paper[19] and expanded in the Humanone model[20] developed by the Copenhagen Institute for Future Studies.

[18] 82 per cent of clinicians say that they do not always have a summary of the care their patients received outside of their practice setting - https://www.infoway-inforoute.ca/en/component/edocman/6407-canadian-interoperability-landscape-study-executive-summary/view-document?Itemid=101
[19] https://mydata.org/publication/mydata-introduction-to-human-centric-use-of-personal-data/
[20] https://cifs.dk/news/the-next-era-in-global-health/

Key principles

1. The first principle is to **ensure that patients are truly at the centre** i.e. offered the possibility to be in charge of their data to ensure high quality, interoperable individual health records. It should enable this extremely valuable and knowledgeable[21] workforce to produce high-quality healthcare data, for the benefit of themselves, their providers, and the healthcare system as a whole.
2. The second principle is to **optimise data reuse for all stakeholders** - in clinical care, clinical research, policy making - through a **'curate once use many times' approach**, where the focus is on curating first data at patient level to generate high quality individual data, before data are aggregated for reuse at population level.
3. The third principle is to go away from the point-to-point data exchange standards and put in place a **data sharing standard** with many-to-many mappings across existing standards, de facto glueing all these standards together, and supporting easy transformation into analytic ready format.
4. The fourth principle is to **use the power of AI technologies to maximise automation in curation** of heterogeneous data; this includes Large Language Model (LLM)[22] and machine learning (ML) based natural language processing to extract information from narrative, ML based entity linking to support coding, ML entity deduplication to identify redundant data, Knowledge Based data quality check.

Emerging solutions

The **MyData Operators Reference model**[23] defines the key functionalities of personal data intermediary organisations supporting citizens to control and manage their data - including the data generated by healthcare organisations, public health, pharma and the patient personally - in a secure environment, ensuring data privacy by design. Pilot initiatives such as the PGO[24] in the Netherlands and the EU CRANE Joint Action[25] demonstrate the value of such Health Data Intermediary (HDI) organisations, allowing citizens to pool all their data and share it with healthcare organisations. There is however one major shortcoming in these emerging initiatives: the data remains in the heterogeneous source format with the limitations identified previously.

The prototype being developed in the **AIDAVA Horizon Europe project**[26] attempts to solve this issue by using multiple curation (AI and non AI-based) technologies to derive (almost) automatically an interoperable personal longitudinal health record from the data pooled

[21] https://link.springer.com/article/10.1057/s41285-024-00208-3
[22] https://ai.nejm.org/doi/full/10.1056/AIe2400548
[23] https://mydata.org/publication/understanding-mydata-operators/
[24] https://www.pgo.nl/
[25] https://crane-pcp.eu/
[26] https://www.aidava.eu

together via a Health Data Intermediary. However AIDAVA-like products can be sustainable at EU level only if following conditions are met.

- There is an (EU wide) agreed reference ontology glueing all existing data standards in use across data sources and functioning as a **Global Data Sharing Standard**.
- All source health data are properly documented in a **FAIR Source Data Catalogue**[27], with technical details on the data schema, metadata, instructions (also called annotations) to transform the source data into the agreed standard ontology through direct mapping for structured data or AI based) transformation tools for unstructured and semi-structured data.

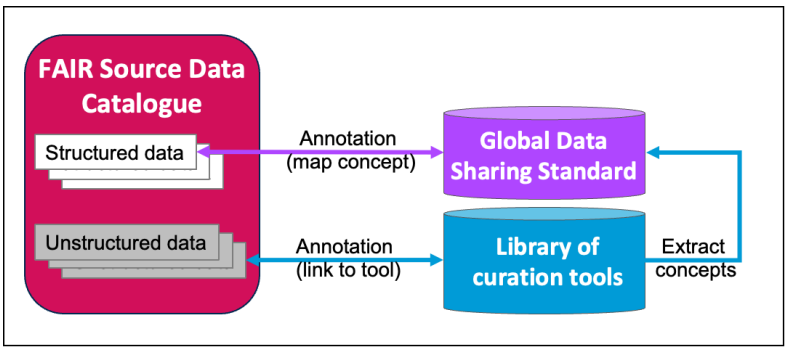

Figure 1. High level view of the FAIR Source Data Catalogue

The AIDAVA prototype is also demonstrating that given the personal health record is interoperable, the implementation of commonly agreed data quality checks (consistency, completeness, etc.) is scalable. This will dramatically increase the quality of the health record and its value in clinical care as well as for secondary use in clinical research and policy-making.

Lessons learned from MyData and AIDVA enable us to derive the following end goal.

*To solve data interoperability in healthcare, **each citizen - or their deputy** - should*

- *be able to **pool their data** through a Health Data Intermediary organisation;*
- *have access to tools enabling ½ **automatically transformation of their data** into a longitudinal health record, compliant with the **global data sharing standard** that supports seamless reuse of data for clinical care, clinical research and policy-making and*
- *be assured of the privacy and security of their data and ability to **control its sharing.***

How would this work in practice

The proposed approach to solve data interoperability is depicted in the figure below.

[27] A Source Data Catalogue is for source data at collection point, EU and Australia are also developing a Dataset catalogue as a repository of reusable secondary data. The technical representation metadata, describing the schema of the data source or data set would be similar; other metadata related to context - such as source system description versus population description - will be different.

- For each data source, a FAIR Source Data Catalogue should be created. This catalogue must include detailed information on the content and mapping to the Global Sharing Standard (i.e. the ontology glueing together multiple standards).
- In compliance with the Data Governance Act[28], all citizens data - including data collected by the citizen through medical devices and patient apps - would be pooled through a private or public Health Data Intermediary organisation selected by the patient.
- The Health Data Intermediary Organisation would provide the citizen an **Intelligent Digital Assistant** to extract the data from each relevant data source and help the citizen to transform and clean their data into an interoperable longitudinal record, compliant with the Global Data Sharing Standard.
- The Health Data Intermediary Organisation would also provide to the citizen a **Digital Wallet** that would include crucial data from the patient such as the International Personal Summary (IPS) and would also help the patient to consent to share data for clinical research and policy making.

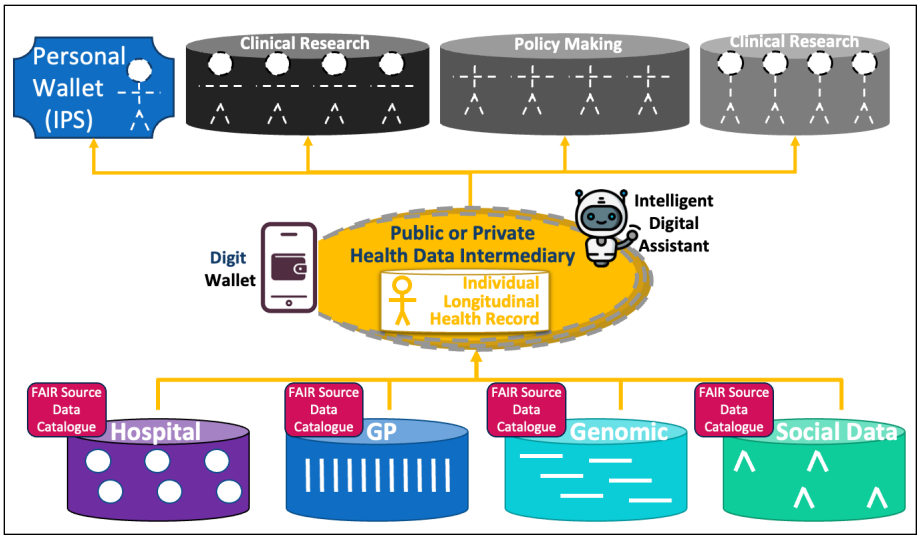

Figure 2. Proposed approach: from data source to reuse

The proposed approach offers a **sustainable and responsible way to solve the health data interoperability problem**, while dramatically increasing the quality of health data for secondary use. The objective is challenging and requires research, proof-of-concept and large-scale feasibility pilots. A structured roadmap as outlined below can achieve this objective.

[28] https://eur-lex.europa.eu/eli/reg/2022/868/oj

Proposed roadmap: 3 steps across the next 10 years

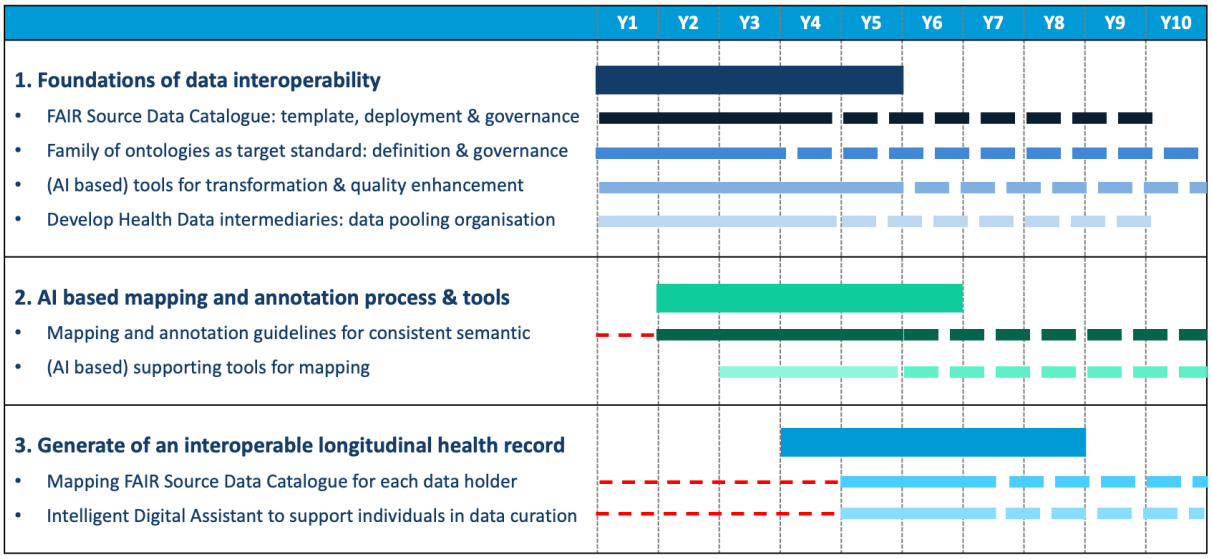

Figure 3. The road to health data interoperability

(the red small dotted lines indicate that research or pilots exists and would enable to consolidate the approach; the larger dotted lines in different colours indicate the need for governance)

Step 1. Establish the foundation of data interoperability (Year1 to Year5)

This step comprises the development of the key components needed to achieve the data interoperability end goal, including establishment of governance processes to ensure long term sustainability of these components.

1. ***FAIR Source Data Catalogue*** with detailed documentation (up to the level of attributes and with context metadata) of all the health systems. Finland - one of the countries with the highest quality data - enforced such an approach in 2013, by law[29].
   The first step is to establish a standard template, like DCAT AP[30] extended with information on technical representation of attributes. Filling the catalogue would require reverse database engineering and could be roughly estimated to require 2,5 person years per system (i.e. ± €500 K); it should be consolidated across vendors[31]
   If supported by the authorities, most data holders should have deployed such a catalogue within 5 years after agreement on the standard template.

2. ***Family of ontologies.*** Agree on a common semantic data sharing standard - glueing together all existing standards used in healthcare such HL7 FHIR, CDISC, openEHR, OMOP, SNOMED, LOINC, MEDDRA, emerging genomic standards within 1M+G

[29] https://aineistokatalogi.fi/catalog
[30] https://www.w3.org/TR/vocab-dcat-3/
[31] Changing a legacy hospital information system - hopefully with good technical documentation - is estimated to cost between €60 M and €100 M for a 1000 beds organisation, i.e. more than 100 times more than delivering such FAIR Source Data Catalogues.

initiatives.. - and building a family of ontologies, linked through a hyper ontology. This should build on the wide experience developed in multiple European projects, including EUCAIM and AIDAVA and more.

The first step is to establish a set of ontological principles and a strict governance process; once agreed, a first hyper ontology could be developed on top of the existing Science Integrated Ontology (SIO)[32] and maintenance could be coordinated by an experienced data standards organisation with in-depth knowledge of semantic and ontology, such as SNOMED.

The resulting ontology will serve as a reference model for the longitudinal health record: all data from the patient will be transformed into an instance of the ontology under the form of a **Personal Health Knowledge Graph (PHKG)[33]** compliant with the ontology.

3. ***(AI) tools for transformation & quality enhancement***. The AIDAVA prototype demonstrates that it is possible to automate - at least partially - the curation process, transforming heterogeneous health data into a single, harmonised PHKG. To maximise automation, it is needed to further develop a set of intelligent tools including
   - digitalisation of paper documents (OCR) ,
   - extraction of structured data from free text in multiple languages (multilingual NLP),
   - medical coding of small piece of free text in multiple languages and terminology alignment (Entity Linking),
   - transformation of structured data into the agreed standard format (Mutate & Transform),
   - managing of duplicated entities and records (Entity Deduplication),
   - data quality check and establishing of data quality label (DQ Validator) ....

   And new tools such as Large Language Model (LLM[34]) should be explored.

   Most tools are based on machine learning models and require ongoing training to improve performance; to ensure they can be deployed in production they will require strict testing and monitoring in compliance with the newly approved AI Act[35].

   Data Quality checks rely on knowledge extracted from consensus documents. They should be further expanded with advanced knowledge extraction methodologies based on LLM[36] to augment verification of consistency across sources. They should also be maintained and re-evaluated on a periodic basis as scientific knowledge evolves and tested for compliance to the AI Act.

   To ensure scalability, the testing, monitoring and retraining approach of these tools will require a structured approach to be defined.

[32] https://jbiomedsem.biomedcentral.com/articles/10.1186/2041-1480-5-14
[33] https://arxiv.org/abs/2104.07587
[34] https://ai.nejm.org/doi/full/10.1056/AIe2400548
[35] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
[36] https://arxiv.org/pdf/2310.06846

4. **Health Data intermediaries (HDI)** are dedicated and certified public or private data organisations that implement data intermediation services as regulated by the EU Data Governance Act[16]. For instance, National Contact Centers For Digital Health (NCDPH) could be considered as emerging public health data intermediaries; private organisations could also be developed in alignment with the PGO in the Netherlands.
   HDIs can serve as trusted partners for patients to control the pooling, integration, curation and quality of their data, and to manage their preferences for sharing their data before it is reused in care delivery, research and policymaking. Data processing with HDIs must be as smooth as possible for the individuals, starting by identifying the health care organisations that have data about themselves that should be pooled within the HDI. Important to note: only the patient - or their deputy - knows where their data can be.
   To meet the first requirement, i.e. data pooling from multiple data sources, HDI must have legally valid data sharing agreements (DSA) with each relevant data holder. Each DSA should include Data Transfer Specification (DTS) aligned with the source catalogue standards template described in point 1. Indeed the content of the DTS should contain the information needed to fill the FAIR Source Data Catalogue.

Step 2. Develop mapping process and (AI) supporting tools (Year2 to Year6)

Semantic mapping is a tedious and difficult process requiring both health and digital/ontology literacy. It is generally accepted that humans may interpret the same concept differently; this leads to non interoperability. During AIDAVA project we observed the same symptom when mapping the attributes of a source catalogue with the reference ontology: one data scientist with medical background but limited experience in knowledge modelling and ontology would map the same concept in a different way that another data scientist with in depth ontology experience but limited health literacy. To maximise interoperability we need to minimise this problem and ensure there is limited/no divergence when mapping data source with the target standard. We therefore suggest developing - and document - a structured process that limits the risk of divergence, while developing AI tools that can replace humans and therefore waive the danger of divergence.

1. **Develop structured process and guidelines on mapping,** inspired from text annotation process and guidelines. To decrease the divergence when annotating free text documents, two approaches are generally recommended.
   - Sequential process: the narrative text is annotated by different people in sequence and inconsistencies are discussed through a well defined process.
   - Parallel process: the narrative text is annotated separately by different people and inter-annotator measures are computed to measure divergence; whenever relevant, an adjudication process allows to come to an agreement across annotators.

An annotation guidelines includes recommendations on the process to follow as well as clear instructions on how to annotate concepts. We suggest establishing an equivalent guideline for the mapping process, ensuring that the annotation/mapping of concepts on attributes defined in source data catalogue is as formal as the annotation of free text used as training datasets of machine learning models. The guidelines should also include quality checks to ensure that it *properly covers all concepts defined in the ontology*.

2. **(AI) tools to support automated mapping**. As FAIR Source Data Catalogues contain existing mapping information and as all catalogues share the same format, it should be possible to train Machine Learning models to perform the mapping automatically. This model should be regularly improved as more annotated catalogues become available.

Step 3. Generate an interoperable longitudinal health record (Year4 to Year8)

The last step is to support individuals - or their deputy - in creating and maintaining their personal longitudinal health record in an interoperable format that is ready for reuse. This would require 2 main steps.

1. **Mapping FAIR Source Data Catalogue of each data holder.** Step 2 described above is expected to deliver the process and tools to map source data with the agreed ontology. This needs to be executed for each data holder, and the resulting mapping can be stored as part of the FAIR Source Data Catalogue.

2. Deploy **Intelligent Virtual Assistants** within the HDI. Most individuals will not be able to manage their data within an HDI unless they have a very intuitive app that supports them. Such a virtual assistant should support the patient in several tasks.
   - Extract the data from all the data holders identified by the patient, following the predefined data sharing agreement between the data holder and the HDI.
   - Integrate the data and curate it into the individual Personal Health Knowledge (PHKG) - referred to in Step 1.2 - compliant with the standard ontology.
   - Check data quality of the PHKG (inconsistencies, errors, incompleteness,..) and provide a data quality label on the overall PHKG, as an indicator for potential data users.
   - Improve data quality. As PHKGs are interoperable, the same tools and models can be applied across all of them. By developing anonymized vector representation of a PHKG, it is possible to develop mathematical models allowing to compare similar PHKGs, for instance citizens with the same genetics and clinical profile, and proactively identify potential gaps and mistakes in the record that can then be corrected by certified health professionals.
   - Publish the six critical categories required by the EHDS regulation (IPS, lab report, discharge summary, prescription and dispensation, medical image report).

- Support patients in sharing data for different purposes, based on smooth dynamic consent management.

A prototype of such an application (apart from the data sharing and consent management part) is developed in the AIDAVA project; the prototype will be further assessed in the summer 2024 in three European hospitals

Conclusion

This document is a first attempt at creating an individuals' centric roadmap to solve the data interoperability problem across healthcare. The proposed approach relies on the extremely valuable workforce that citizens represent.

This would benefit further discussion. The facts however remain: we need a structured roadmap if we want to solve this ever growing problem. With the increasing amount of data flowing into the health systems and the potential of data greedy AI technologies this becomes more important every day.

As demonstrated in this document, the individual patients - or their deputy - are at the core of the solution as active stakeholders:

- only patients know where all their personal data are,
- only patients can add the personal data they are collecting,
- only patients - and their healthcare providers - are interested to have a interoperable and reusable personal dossier,
- only patients - and potentially their healthcare providers - can identify errors in their health record.

Collective data interoperability can be achieved by ensuring that each patient's data is available in the agreed interoperable format.

To solve the data interoperability issue we need a structured roadmap.

As demonstrated in this document, the individual patients - or their deputy - are at the core of the solution as active stakeholders.
Patients so far have been considered as passive stakeholders in all aspects of health data management; their roles should be reconsidered and they should be treated as a core player in the overall solution.

Acknowledgements

The ideas described in this paper were inspired by the lessons learned from implementing the first generation of the **AIDAVA** prototype, an AI based digital virtual assistant aimed at supporting citizens to curate their data. AIDAVA is developed under Horizon Europe Grant Agreement no: 101057062. The concept of Health Data Intermediaries is directly inspired by

the work of the **MyData Health Community** around a human -centric digital health ecosystem.

Prof Dipak Kalra from The European Institute for Innovation through Health Data, Belgium and Med.Dr Lars Lindsköld, Scilife Lab datacenter, Sweden provided insightful comments to the document.