Navigating the Health Data Wilderness in the Dawn of the Data Age

Data Value Chain White Paper

Navigating the Health Data Wilderness in the Dawn of the Data Age

Data Value Chain White Paper

Zoltán Lantos[1], Kate Spyby[2], Peija Haaramo[3], Mari Mäkinen[4], Péter Fésüs[5], Rosalyn Moran[6], Robert Verheij[7], Truls Korsgaard[8], Anne Heidi Skogholt[8], Thomas Kvist[9], Fredric Landqvist[10]

The work has been carried out in the Metadata Subgroup of SITRA's National Initiatives Network

1 ___ Department of Virtual Health Guide Methodology, Semmelweis University Faculty of Health Sciences, Hungary
2 ___ Australian Institute of Health and Welfare (AIHW), Australia
3 ___ Finnish Social and Health Data Permit Authority Findata, Finland
4 ___ Finnish Institute for Health and Welfare (THL), Finland
5 ___ eHealth Software Solutions Ltd, Hungary
6 ___ EKOS - Social and Environmental Research Associates, Ireland
7 ___ Nivel, Netherlands
8 ___ Directorate of e-health, Norway
9 ___ Region Västerbotten, Sweden
10___ Tietoevry AB, Sweden

Reviewers
Cátia Sousa Pinto, Shared Services for Ministry of Health, Portugal
Gábor Bella, IMT Atlantique, France

November 2023
Department of Virtual Health Guide Methodology, Semmelweis University Faculty of Health Sciences

TABLE OF CONTENT

ABBREVIATIONS

AI	Artificial Intelligence
AIHW	Australian Institute of Health and Welfare
API	Application Programming Interface
ATC	Anatomical Therapeutic Chemical Classification System
BERT	Bidirectional Encoder Representations from Transformers
CDA	Clinical Document Architecture
CLL	Chronic Lymphocytic Leukemia
CSV	Comma-Separated Values
DCAT-AP	Data Catalog Vocabulary - Application Profile
DDI-L	Data Documentation Initiative - Lifecycle
DIGG	Swedish Agency for Digital Government
DLU	Data Linkage Unit
DPR	Dense Passage Retrieval
DQS	Data Quality Statements
DSS	Data Set Specifications
EDQM	European Directorate for the Quality of Medicines and HealthCare
EHRs	Electronic Health Records
ETL	Extract-Transform-Load
FAIR	Findable, Accessible, Interoperable, and Reusable
FEAM	Federation of European Academies of Medicine
FHIR	Fast Healthcare Interoperability Resources
GDPR	General Data Protection Regulation
GP	General Practitioner
GPS	Global Positioning System
GPT-3	Generative Pre-trained Transformer 3
GRADE	Grading of Recommendations, Assessment, Development, and Evaluations
GSIM	Generic Statistical Information Model
HIPAA	Health Insurance Portability and Accountability Act
HL	Hodgkin Lymphoma
HL7	Health Level Seven International
IC	Informed consent
ICD	International Classification of Diseases
IoT	Internet of Things

ISO/IEC 11179	International Organization for Standardization and the International Electrotechnical Commission Information technology — Metadata registries
JSON	JavaScript Object Notation
LOINC	Logical Observation Identifiers Names and Codes
mCODE	Minimal Common Oncology Data Elements
METEOR	Australian Government Online Metadata Registry
ML	Myeloid Leukemia
MM	Multiple Myeloma
NABR	National Affective Diseases Registry
NBEDS	National Best Endeavours Data Set
NBPDS	National Best Practice Data Set
NFBR	National Infectious Diseases Registry
NGO	Non-Governmental Organisation
NHBR	National Haematological Diseases Registry
NHL	Non-Hodgkin Lymphoma
NMDS	National Minimum Data Set
NSR	National Stroke Registry
OMOP-CDM	Observational Medical Outcomes Partnership Common Data Model
PCR	Polymerase Chain Reaction
PREM	Patient Reported Experience Measures
PROM	Patient Reported Outcome Measures
PROV-O	Provenance Ontology
RAG	Retrieval Augmented Generation
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SHA-256	Secure Hash Algorithm 256-bit
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
THL	Finnish Institute for Health and Welfare
TTP	Trusted Third Party
UCUM	Unified Code for Units of Measure
UNICOM	Universal Decimal Classification
WHO	World Health Organization

FOREWORD

The transformation of healthcare is deeply interwoven with the proliferation of data. However, a fundamental truth remains: data, if merely collected and not utilised, is associated with costs and risks but offers no intrinsic value. Its true power is revealed when applied effectively to improve health outcomes and enable precise decision-making in a timely manner, ensuring that we are consistently "doing the right things right."

As we journey through the Data Age for Health & Care, it becomes evident that data is not just abundant, but a renewable resource, serving as the cornerstone of value creation. This paper shines a light on the intricate data value chain, viewing individuals as data centres. It also delves into the innovative mechanisms of the data loop, posited as a potential European platform model, while addressing the challenges faced in data analysis.

Data is not just transforming the healthcare landscape; its ripple effect is visible across society, marking its place as a new type of valuable asset. Central to our discourse is the transition of health-related raw data from mere reality to a structured data space.

Furthermore, this paper takes a deep dive into the mechanics of converting data into value, scrutinising the ethics and legislations that envelop it, and the larger economic implications of data utilisation. This paper also touches upon essential facets of data architecture, including but not limited to metadata, data quality, and integration. The challenges posed by data anomalies, from outliers to systematic errors, are also meticulously addressed.

By highlighting best practices in the domain — ranging from the metadata registry to Swedish API advancements — and shedding light on innovative research topics, this paper provides a comprehensive panorama of the data landscape in health and care. The underlying message is unequivocal: data, when leveraged correctly, is instrumental in enhancing patient care, fostering value, and driving innovation.

As you delve deeper, this paper will not only offer insights but will also provoke reflection on our collective journey towards a data-driven health ecosystem. Dive in and discover the transformative potential of data in shaping a healthier, informed, and innovative future.

Bogi Eliasen and Aron Szpisjak, Copenhagen Institute for Future Studies, Denmark

MANAGEMENT SUMMARY

This paper discusses the foundations of the Data Age for Health & Care, and how data can be transformed into value. It highlights the key characteristics of the data value chain, including data as a renewable resource and the individual as the data centre, examines the mechanism of the data loop as a potential new European platform model and the limitations and hindrances of data analysis. Good practices, such as metadata registry and improved data utilisation, are also discussed. Research and innovation topics are proposed, including the valuation of data and the development of standards for documentation and recording.

The significance of data in transforming society is explained, and how it can be considered a renewable resource and a new type of good. It discusses health-related raw data and what is needed to transform it from reality into data space.

A separate section focuses on how data can be converted into value, highlighting the mechanism of a data loop, value-based platform model, and the ethics and legislation surrounding it. The economy of data is also discussed, both in terms of what we already know and what we need to find out.

The paper then moves on to discuss how individuals are becoming data centres, which model is known as the Humanome. It covers elements of data definition and architecture, such as metadata, data quality, standard formats, knowledge graph, and data integration. Limitations and hindrances such as incomplete data flow, outliers, incorrect data, systematic errors, and small datasets are also explored.

Good practices are presented, such as metadata registry, improved data utilisation, regulation, quality patient registry, Swedish API efforts, and Nordic Common Variables in the metadata catalogue. Value modelling and estimations are discussed, and research and innovation topics are proposed, such as domain-specific descriptions of data mesh linked with semantic mesh, documentation and recording standards for coded versus free text input, documentation and recording standards for the two types of knowledge of health professionals – i.e. the evidence-based literature knowledge and the experience based tacit knowledge, and data valuation studies.

Overall, our paper highlights the importance of data in the health & care domain and how it can be utilised effectively to provide value, improve patient care, and lead to innovation.

1. THE FOUNDATIONS OF THE DATA AGE FOR HEALTH & CARE

1.1 DATA TRANSFORMING OUR SOCIETY

Data has already transformed our society in numerous ways, shaping how we live, work, communicate, and make decisions. The availability and analysis of data have revolutionised various aspects of our society, leading to significant changes.
Data-driven decision making has become a core element of modern **businesses**. Companies use data to identify customer preferences, optimise operations, and develop new products and services. Data-driven insights also enable businesses to tailor their marketing strategies, target specific demographics, and personalise customer experiences. Additionally, data has facilitated the rise of new business models, such as data-driven start-ups and the sharing economy.

In the **healthcare** industry, data-driven solutions enable better patient care, diagnostics, and treatment. Electronic health records (EHRs) and health data analytics have improved patient outcomes, enabled personalised medicine, and facilitated research and development of new drugs and therapies. Data also plays a critical role in public health, helping to identify disease outbreaks, track health trends, and inform policy decisions.

Educational institutions use data to personalise learning experiences, track student performance, and identify areas where additional support is needed. Data-driven learning platforms provide personalised feedback, adapt to individual learning styles, and offer customised content. Data also enables online and remote learning, making education more accessible to a wider population.

Data has changed the landscape of **governance** and **public policy**, enabling evidence-based decision making. Governments use data to inform policies on a wide range of issues, from urban planning and transportation to public health and social welfare. Data-driven technologies, such as smart cities, use data to optimise resource allocation, improve public services, and enhance citizen engagement.

Data has transformed **social** and **cultural** dynamics, shaping how we communicate, interact, and perceive the world. Social media platforms generate massive amounts of data that influence public opinion, shape social trends, and drive cultural change. Data-driven algorithms also impact our online experiences, from personalised content recommendations to targeted advertising. Data has also brought about new forms of artistic expression, such as data visualisation and digital art.

Data will continue to play a transformative role in our society, here are some potential ways data may further transform our society in the future.

As data continues to accumulate, advancements in **artificial intelligence (AI), automation**, and the development of **new models** such as large language models are expected to accelerate.

Data-driven AI systems can analyse vast amounts of data to identify patterns, make predictions, and automate tasks across various industries. This could lead to increased efficiency and productivity in sectors such as manufacturing, logistics and also healthcare, but may also impact the workforce, requiring reskilling and adaptation to changing job roles.

Data-driven **personalisation** is likely to become even more prevalent in the future. With the increasing availability of data from various sources, including wearable devices, social media, and the Internet of Things (IoT), personalised experiences could become even more tailored to individual preferences and behaviours. This could include personalised health plans, customised products and services, and hyper-personalised communication, among others.

Data will continue to play a pivotal role in the development of **smart and healthy cities**, where interconnected devices and sensors generate vast amounts of data to optimise urban planning, resource allocation, and citizen services. The IoT will enable smart infrastructure, such as smart grids, smart transportation, and smart buildings, which can provide real-time data for better decision making and resource management.

Health & care is expected to see significant advancements through data-driven innovations. Precision medicine, which uses genetic, environmental, and lifestyle data to tailor treatment plans, is likely to become more widespread. Telemedicine, remote monitoring, and digital health solutions could leverage data to provide better access to health & care services and improve patient outcomes.

As data continues to play a pivotal role in society, ethical considerations and data governance will become even more crucial. Ensuring responsible data collection, storage, and use, protecting individuals' privacy rights, addressing issues of bias and fairness in data-driven technologies, and developing robust ethical frameworks for data use will be critical to ensure that data continues to be leveraged for positive societal impact.

Data-driven decision making is expected to become even more prevalent in governance and public policy. Governments may use data to address complex societal challenges, such as climate change, poverty, and inequality, and to develop evidence-based policies. Open data initiatives may gain further momentum, fostering transparency, accountability, and citizen participation in decision making.

Data-driven entrepreneurship could also be shaped by ethical considerations, with the rise of ethical data start-ups and social enterprises that prioritise responsible data use and societal impact. **Ethical data entrepreneurship** may involve leveraging data for social good, addressing social and environmental challenges, and promoting fair and equitable data practices.

In conclusion, data will continue to transform our society in various ways, with advancements in AI, personalised experiences, smart cities, health & care innovations, data governance, governance decision making, and ethical data entrepreneurship. Emphasising responsible data use, privacy, fairness, and ethical considerations will be critical to ensure that data continues to drive positive societal transformation.

1.2 DATA AS A RENEWABLE RESOURCE

Data can be considered as a renewable resource due to several key reasons.

It has the ability to replenish, renew and even enrich itself. With the advancement of technology, data is constantly being generated and collected from various sources, including electronic health records, digital devices, sensors, social media, location data, environmental data, and more. This continuous generation of data ensures a constant supply of new information and, what is more, potential refinement and enrichment, making data a renewable resource.

Data can be easily replicated and shared across multiple platforms and systems without depleting the original source. This means that data can be copied, stored, and distributed across different applications and locations, enabling multiple users to access and utilise it simultaneously. This characteristic of data allows for its efficient use and scalability, making it akin to a renewable resource.

It can be transformed, processed, and repurposed into different formats and structures for various applications. Data can be analysed, aggregated, and used for different purposes, such as business intelligence, scientific research, and decision-making. This recyclability of data ensures that it can be repurposed and reused in different contexts, just like how renewable resources can be utilised in various ways.

Data, when managed and stored properly, can have a long lifespan and can be used for extended periods of time. Data can be preserved through data backup and archiving methods, ensuring its availability for future generations. Additionally, data can be curated to maintain its accuracy, relevance, and usefulness over time, making it a sustainable resource that can be utilised for the foreseeable future.

Unlike traditional resources such as fossil fuels, data itself does not have a physical presence and does not require extraction or consumption of natural resources. This may reduce the environmental impact associated with data generation and usage, potentially making it an eco-friendly resource that can with care be sustainably utilised without depleting natural resources. However, the current technology, with its ever-growing computational demand, has very high energy consumption, and thus the promise of being eco-friendly is not fulfilled yet.

Overall, data can be considered as a renewable resource due to its ability to regenerate, replicate, be recycled, sustainably managed, and reduce environmental impact. As technology continues to advance, the availability and potential uses of data are expected to increase, further solidifying its status as a renewable resource.

1.3 DATA AS A NEW TYPE OF GOOD

Data has emerged as a distinct category of goods, with unique characteristics and behaviour that differ from traditional goods and resources. As we navigate the era of digital transformation that is leading us into the Data Age, data is fundamentally changing our society and economy in ways we cannot fully anticipate.

Data, as a resource, is derived from the digital footprint, which is a digital replica of the real world. However, the accuracy of this replica is often not rigorously tested or evaluated in terms of projection methodology. This variability in quality ranges from high accuracy in digital medical imaging to low accuracy in mapping mental diseases and ill health, representing extremes on the quality scale.

The optimal digital space serves as a multidimensional replica of the real world, constructed through a precisely defined projection methodology, similar to cartography. The metadata associated with data points defines their exact location in this multidimensional data space.

As we continue to grapple with the transformative power of data, recognising its unique characteristics as a new type of good is crucial for navigating the digital age effectively.

1.4 HEALTH RELATED RAW DATA

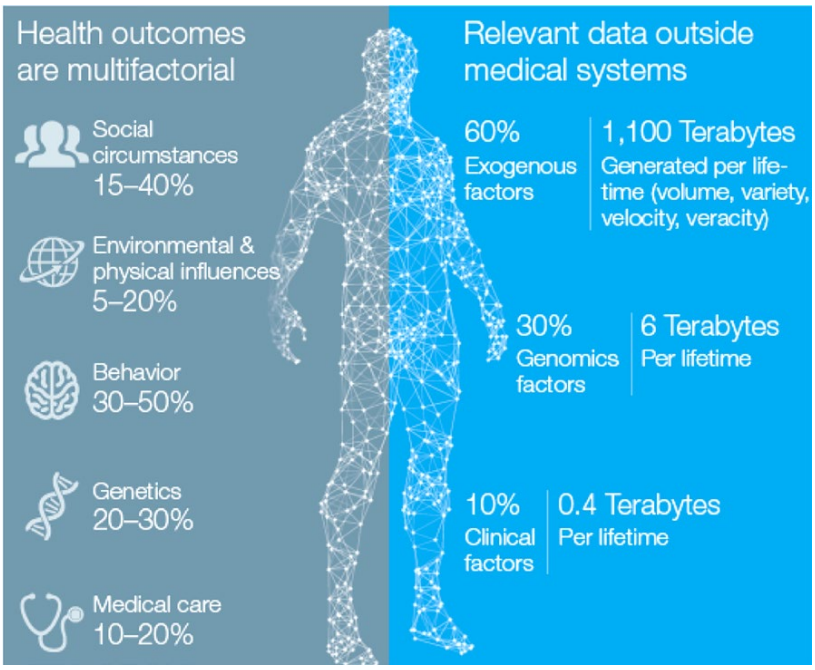


Figure 1: The Relative Contribution of Multiple Determinants to Health Outcomes
Source: McKinsey & Company, 2017 based on Health Affairs, 2014 [1]

Raw data is now mostly seen as a resource that will revolutionise medicine and health, as the FEAM (Federation of European Academies of Medicine) Forum Annual Lecture 2022: Digital Health and AI on 26 October (https://www.youtube.com/watch?v=c3dw4ImoUNc)[2] recently demonstrated.

[1] Health Policy Brief: The Relative Contribution of Multiple Determinants to Health Outcomes, Health Affairs, August 21, 2014.
[2] FEAM Forum Annual Lecture 2022: Digital Health and AI on 26 October, https://www.youtube.com/watch?v=c3dw4ImoUNc

However, data must be processed to be used as a valuable and reliable resource. In contrast to data mining, precise data definition is needed.

Precise data definition is the foundation for effective data processing. It involves defining and describing data elements with accuracy, clarity, and consistency. Key aspects of precise data definition include:

- Data Modelling: Data is represented using formal models, such as entity-relationship diagrams, to define the structure and relationships between data entities.
- Data Standardisation: Establishing consistent data formats, naming conventions, and data types helps maintain data quality and interoperability.
- Metadata: Metadata provides context and descriptions for data elements, making it easier to understand and use the data.
- Data Documentation: Comprehensive documentation includes data dictionaries and data catalogues that describe the meaning, source, and usage of each data element.
- Data Governance: Data governance policies and practices ensure that data is defined and ma aged consistently across the organisation.
- Data Trajectory: The process of capturing, recording, and representing real-world events, objects, or phenomena as data points or information in a digital format.

1.5 TRAJECTORY – FROM REALITY TO DATA SPACE

Datasets, and data space more generally, are most valuable when they enable us to perform precise operations on them that contribute to learning about reality, most usefully in the form of predictions. The most valuable data space is thus the most accurate digital copy of reality possible.

Data as a resource is a methodologically defined projection of reality. Data point is the point of reality mapped into virtual space with a precise definition. Today, accuracy can generally be defined according to the purpose of use, but for systematic scientific study, it is desirable to achieve uniform projection accuracy. By defining the data in a methodologically clear way, a data space can be built, the accuracy and quality of which can be determined by methods to be further developed and elaborated.

Data science is able to perform value-creating operations in a virtual space, which is built from properly defined data points as a projection of reality.
Motto: *"Performing data science operations in an accurate data space is like playing a fine instrument."*

The level of projection of reality determines the value of data as a resource, and consequently the validity of the results of data science operations. However, the definition of the level of projection has not yet been properly elaborated.

The concept of fit for purpose is used, which only applies to known segments of the virtual space. In the majority of the cases, the true fit is not known in health, and neither is the exact validity of the results. A more accurate and comprehensive data definition will improve the fit.

From a clinical perspective, "...most diseases are umbrella terms lumping together different causal mechanisms that share this one name-giving phenotype. Therefore, the use of such umbrella disease terms in biomedical research and clinical practice generates an impenetrable mix of molecular mechanism and clinical comorbidities. (...) Since genetics and clinical medicine seem to live in separate classification systems, taxonomies, or ontologies, the ever-increasing wealth of genetic information does not lead to innovation."[3]

A systems approach is able to link data science practice, homeostasis as a system and systems medicine.

A possible good systems data architecture follows the general biobank structure.
Biomarkers provide the basic dataset, by measurable characteristics, such as genetic or biochemical markers, that can indicate a specific biological condition or disease. Biobanks may collect and store various types of biomarkers, such as DNA, RNA, proteins, metabolites, and other molecules, from individuals or populations.

Samples: Biobanks collect and store biological samples, such as blood, tissue, urine, saliva, or other types of specimens, that contain biomarkers of interest. These samples are collected from individuals or populations and stored in a controlled environment, typically at low temperatures, to preserve their integrity and quality for future research.

Individuals: Biobanks may collect samples and associated data from individual donors, who may be patients with specific medical conditions, healthy individuals, or participants in research studies. The individuals may provide informed consent for their samples and data to be used in research, and their privacy and confidentiality are protected in accordance with ethical and legal requirements.

Cohorts: Biobanks may also establish cohorts, which are groups of individuals who share certain characteristics or exposures, and who are followed over time to study health outcomes or other research questions. Cohorts may be established based on specific criteria, such as age, gender, ethnicity, geographical location, or specific health conditions, and may involve long-term follow-up to gather data on health outcomes, lifestyle factors, and other relevant information.

Entire Population: Some biobanks may aim to collect samples and data from an entire population, such as a specific geographic region, a defined population group, or a particular demographic. These population-based biobanks can provide valuable resources for understanding health trends, disease prevalence, genetic variation, and other population-level characteristics.

Good examples of well-established projection methodology between reality and data are the calibration techniques and attribution methods of laboratory tests, genetic profiling, PCR (polymerase chain reaction), and reverse transcription polymerase chain reaction (RT-PCR). Just like complemented image projection technologies, such as annotated medical imaging.

[3] Zeinab M. Mamdouh, Elisa Anastasi and Ahmed A. Hassan et al. Why the way we define diseases prevents innovation and precision medicine. DrugRxiv. 2023. DOI: 10.14293/S2199-1006.1.SOR-.PPCFYDY.v1

As location and the community living there are the most important determinants of health, the most viable accurate projection is anchored by spatial location.

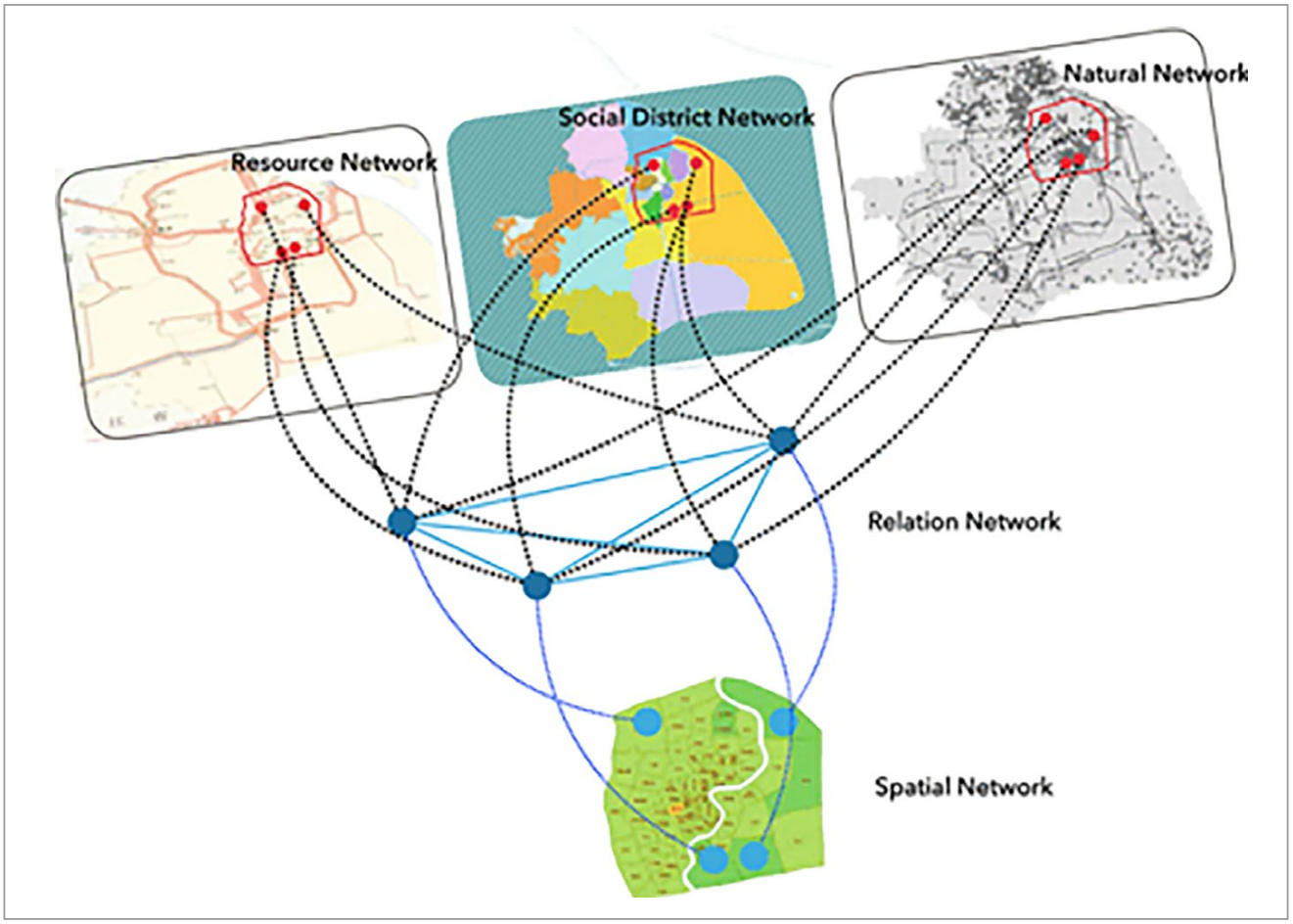


Figure 2: Spatial location as a projection basis[4]

We should also take into account recent developments of the spatial web https://medium.com/swlh/an-introduction-to-the-spatial-web-bb8127f9ac45, which bring us closer to a useful multidimensional data space linked to our spatial living space.

To enable the dynamic study and management of the data footprint, time is also an inherent attribute in data projection of reality. Thus a timestamp is an inevitable element of data description – of the provenance section of metadata. This way, provenance, as a key metadata element from which trust in data accuracy is derived, becomes enriched with an additional valuable dimension.
The interconnected and hierarchical network of systems conceptualised here allows us to describe health and diseases more accurately than we do now, and provides an unparalleled opportunity to observe them by data science methods applied in mathematics and physics.

[4] Bai et al: STG2Seq: Spatial-temporal Graph to Sequence Model for Multi-step Passenger Demand Forecasting, 2019. https://doi.org/10.48550/arXiv.1905.10069

2. FROM DATA TO VALUE

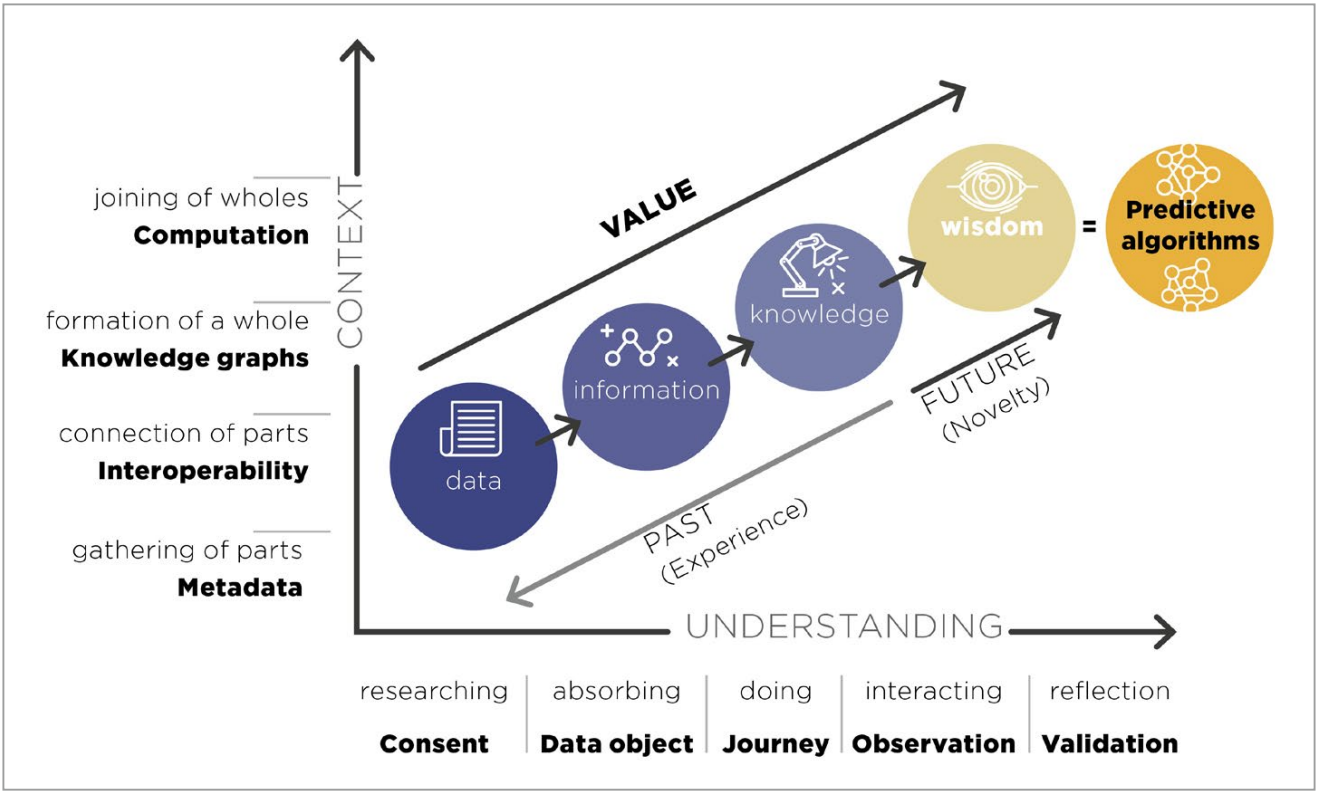


Figure 3: The Data to Prediction Value Chain[5]

The Data to Prediction Value Chain (Figure 3) unfolds a systematic approach to data management, analysis, and interpretation in the context of health sciences. Axis Y navigates the intricate process of assembling disparate elements into a cohesive structure, emphasising the importance of data definition and the methodology of projection. The narrative extends to the establishment of a comprehensive knowledge graph, detailing the encoding of individual health situations to overarching health targets, fostering meaning-level interoperability. Meanwhile, Axis X delves into the research continuum, underscoring the significance of ethical data collection methodologies, with a spotlight on the burgeoning role of IoT and the centrality of data consent. The subsequent phases involve the absorption of data into the right frameworks, elucidating processes through digital storytelling, and interactive exploration of patterns via data science. The process culminates in a reflective phase, emphasising validation through statistical analysis and clinical trials, completing a holistic framework for data-driven wisdom in health sciences.

AXIS Y
gathering of parts – data definition
⟶ methodology of projection *(see 5.1. metadata)*
connect parts – data network through system-level interoperability
⟶ standards for data storage and operations *(see 5.3. standards)*
formation of a whole – knowledge graphs from individual health situation to health target
⟶ standardised encoding of situation specific health (patient) journey variations *(see 5.4. knowledge graph)* ensuring meaning-level interoperability
joining of wholes – patterns, regularities of knowledge graphs ⟶ predictions
⟶ computation – data science, machine learning, artificial intelligence

A possible analogy for the role of metadata is sheet music, in which we can see little data, rather a standard representation.

[5] Lantos 2022, based on Ackoff 1989 in: The Next Era in Global Health by Copenhagen Institute for Futures Studies, 2020.

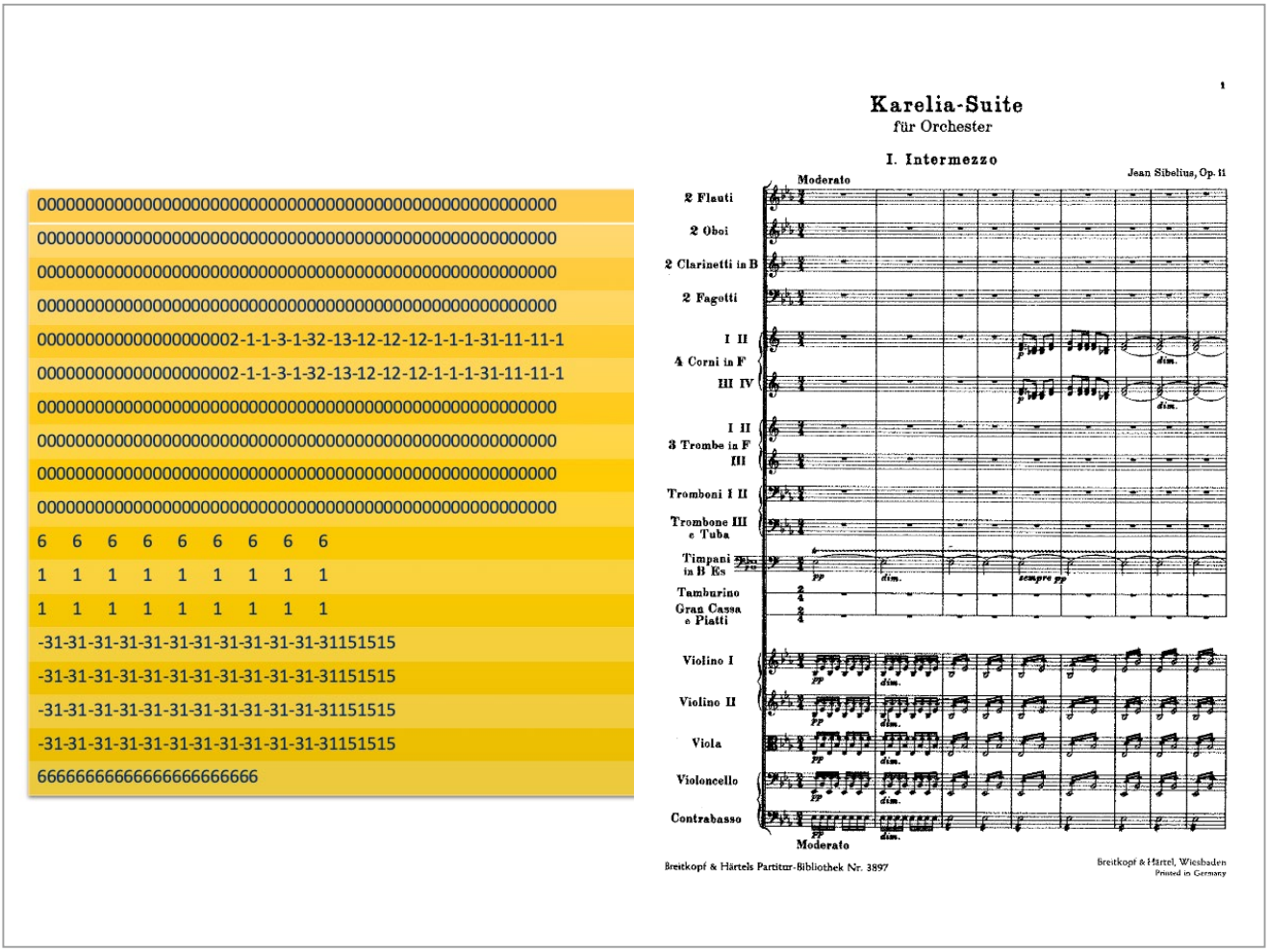


Figure 4: From data to metadata – the initial dataset and sheet music of Sibelius Karelia Suite

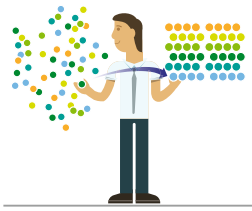
Or we now use GPS coordinates for location, however a map uses conventional symbols. For AI, an interesting analogy from Luciano Floridi (Universities of Oxford and Bologna) that the current practice sees, and consequently regulates, AI as a product – a microwave oven – instead of a procedure like cooking.
https://www.linkedin.com/feed/update/urn:li:activity:6993933927417593856/?utm_source=share&utm_medium=member_desktop
Likewise, the same applies to data, where we differentiate between primary and secondary data usage. However, there is no distinction in terms of data discovery, accessibility, interoperability, and reusability, whether we're extracting a dataset, such as diastolic blood pressure over a specific time frame, to understand an individual's medical history or to develop an algorithm for personalised prediction in preventing high diastolic blood pressure.
Rather, it is worth distinguishing between the process of using the data for algorithms in operation or to develop new algorithms in research, development and innovation.

Our proposed value chain approach is a sequence of procedures all the way through.

AXIS X

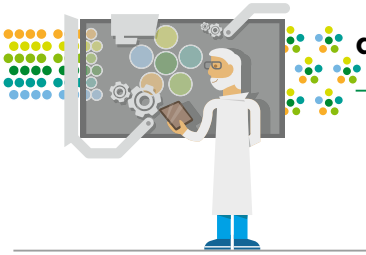
researching – data obtained

→ method of data collection,
role of IoT is growing + data consent is key



absorbing – putting it in the right place/data definition

→ architecture, format of data objects

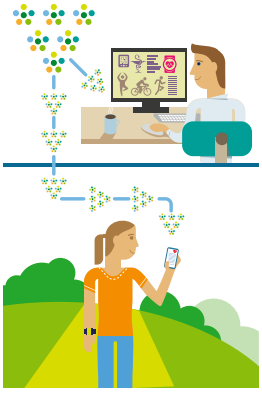


doing – describing processes/journeys/"stories"

→ digital story telling – branched network of activity sequences for example from an event catalogue, with input data, activity status data and output data

interacting – exploring patterns, describing phenomena

→ observation through data science – mostly data-based analysis, less model-based analysis



reflection – validation

→ statistical analysis comparing the prediction with the real phenomenon, clinical trials

Most of the financial resources and attention are currently devoted to interoperability and machine learning (connection of parts/reflection).
Insufficient resources are devoted to data definition and knowledge graphs, – as well as data science as an observation methodology (formation of a whole/interacting).

For the best practices in the data-to-prediction value chain, a multidisciplinary approach that couples data governance with knowledge governance and data stewardship appears to be viable.

Data governance focuses on managing the availability, usability, integrity, and security of the data used by any kind of organisation (single entity, a project organisation or virtual organisation). In order to effectively manage the data, it is important to understand how it is being used and what the organisation's goals and priorities are with regards to that data. Knowledge governance provides this context by ensuring that the data is used in a way that supports the organisation's goals and aligns with its values and policies.

Without knowledge governance, data governance may struggle to ensure that the data is being used effectively and efficiently. For example, if data quality is not managed in a way that supports the organisation's goals, the data may be of poor quality and not suitable for decision-making. Similarly, if the data is not properly secured, it may be vulnerable to unauthorised access or misuse.

In the context of health & care, knowledge governance is linked to the quality of the evidence on which recommendations are based, so knowledge graphs can be composed of sections representing different levels of quality.

In accordance with the systemic approach to health previously presented (page ...), it is worth introducing a new quality attribute called "fidelity." This attribute describes the maximum level of correspondence to the dynamic processes of the human organism's reality, as defined by the quality of scientific evidence, with which we can describe the specific part of the system in question. Quality of evidence is evaluated by the GRADE system.[6]

[6] GRADE = Grading of Recommendations Assessment, Development, and Evaluation (https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade/)

3. MECHANISM – DATA LOOP

3.1 VALUE-BASED PLATFORM MODEL

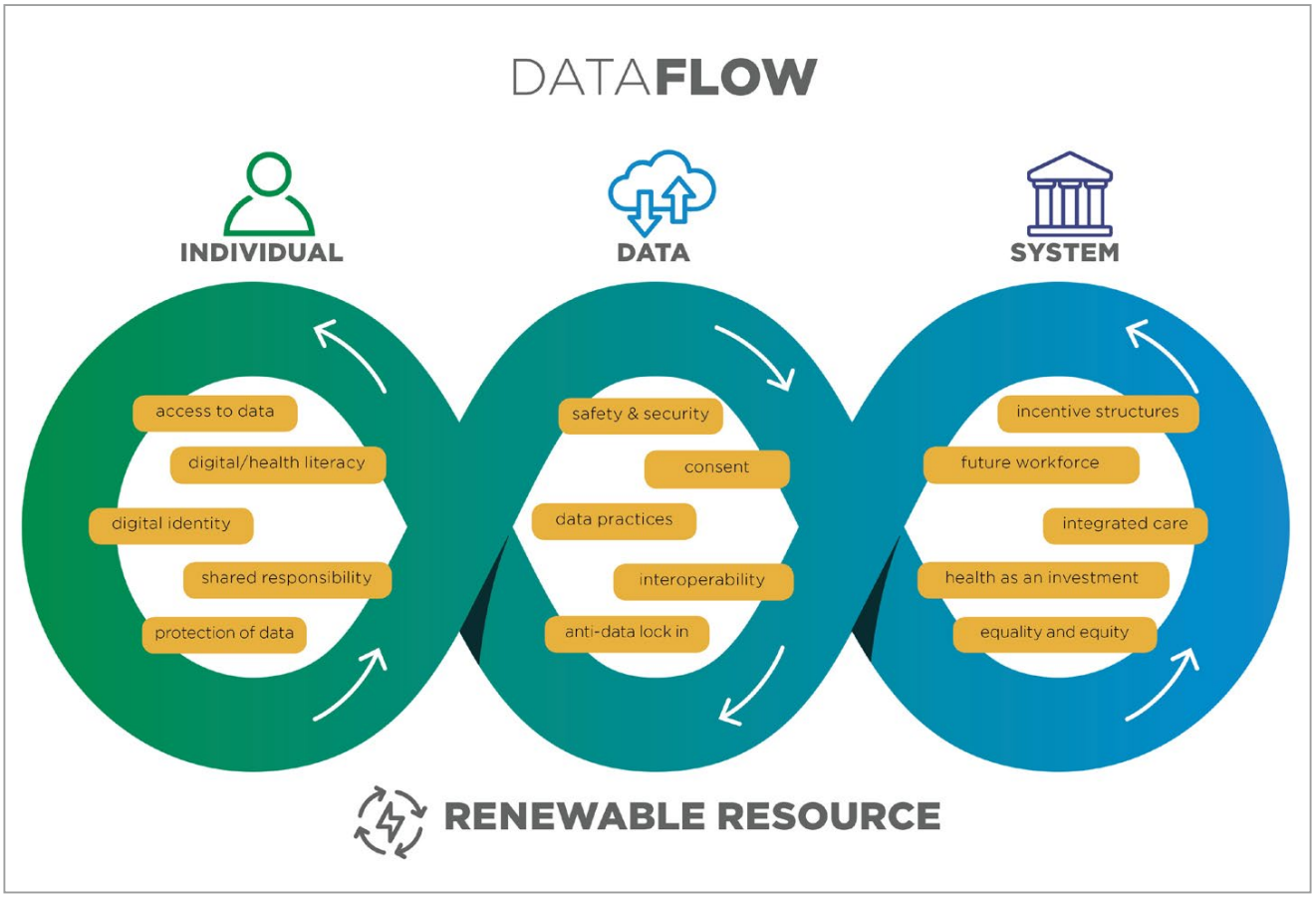


Figure 5: Flow in the data loop for high utilisation[7]

INDIVIDUAL: Data donation – Data: process & use – System: Organisations providing benefit
DATA DONATION: access to data, data/health literacy, shared responsibility, protection of data, digital identity.
DATA USAGE: safety & security, consent, interoperability, anti-data lock-in, data practices.
Organisations providing benefit: incentive structure, health as investment, equality & equity, integrated care, future workforce.

Institutes and national/regional systems are important mediators and distorters of individuals' interest. (In the new European Data Governance Act, data intermediation service providers are regulated as a separate legal entity.)

CONFLICT: there is no clear evidence that widely shared individual data can create significant health value, so stakeholders are reluctant to share data, except chronically and/or seriously ill patients, despite clear data sharing principles.

[7] The Next Era in Global Health by Copenhagen Institute for Future Studies, 2020.

The reluctance is very much understandable, considering:

- the vulnerability of data security and privacy mechanisms in the real world;
- the significant market pressure to commoditise the health of individuals and their data for profit;
- the absence of evidence or examples, as mentioned earlier, that the individual can personally benefit from sharing their data.

In the context of data, it can be seen as a renewable resource when there is a constant flow of data. However, to effectively harness the potential of data as a renewable resource, a health-specific platform model is needed to facilitate the flow of data in a secure and regulated manner.

The benefits to individuals from donating their data are crucial, as they can serve as the engine that drives the flow of data. To incentivise individuals to donate their data, alternative remuneration and incentive models are necessary, such as those based on Digital Ledger Technology or Blockchain. These innovative models can provide individuals with rewards and compensation for their data contributions, creating a more sustainable and mutually beneficial ecosystem for data sharing and utilisation. By exploring and implementing such novel approaches, we can unlock the full potential of data as a renewable resource and foster responsible and ethical data practices in the digital age.

The apparent contradiction between privacy and FAIR (Findable, Accessible, Interoperable, and Reusable) data can be effectively resolved by considering the use case-specific usefulness and the overall utility of the health-specific (European) data platform. It is important to strike a balance between protecting privacy and promoting data accessibility and usability in a way that maximises benefits for all stakeholders. While market mechanisms may not always be the most appropriate tools for determining the value of data and ensuring the best outcomes for citizens, recognising data as a public good can be a major driver of health and care in all European Member States. This perspective aligns with the European Value Set , which places emphasis on self-transcendence and the greater good. By prioritising the collective benefit of data sharing and utilisation in the context of health and care , we can navigate the complexities of privacy and data utility while upholding the values that promote the well-being of individuals and societies as a whole.

[8] Witte EH, Stanciu A, Boehnke K: A New Empirical Approach to Intercultural Comparisons of Value Preferences Based on Schwartz's Theory. Front. Psychol. 11:1723. 2020

[9] Prainsack, B et al.: Data solidarity: a blueprint for governing health futures. The Lancet Digital Health, Volume 4, Issue 11, e773-e774. 2022

3.2 ETHICS AND LEGISLATION

Ethics and legislation related to data in health & care are crucial to ensure responsible and ethical use of data to protect individuals' privacy, promote fairness, and foster trust in the healthcare system.

Health & care data often contains sensitive and private information about individuals' health conditions, treatments, and personal identifiers. Protecting the **privacy and confidentiality** of this data is paramount. Ethical considerations and legislation, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States, The Privacy Act 1988 (Privacy Act) in Australia, and the General Data Protection Regulation (GDPR) in the European Union, mandate strict safeguards for the collection, use, and disclosure of health & care data, including consent requirements, data encryption, and data breach notifications.

Obtaining informed **consent** from patients or research participants is an important ethical principle in health & care data collection and research. Informed consent ensures that individuals are adequately informed about the purpose, risks, benefits, and potential uses of their data and have the right to control how their data is used. Ethical guidelines and legislation often require obtaining explicit and informed consent for data collection, storage, and use, and may also require consent for secondary uses of data beyond the original purpose of data collection.

Ensuring the **security and integrity** of health & care data is vital to protect against data breaches, unauthorised access, and data manipulation. Ethical considerations and legislation may require healthcare organisations to implement robust data security measures, including encryption, access controls, and audit trails, to safeguard against data breaches and protect the integrity and confidentiality of health & care data.

Ethical considerations and legislation related to **data sharing and collaboration** in health & care are critical to facilitate research, innovation, and evidence-based decision making. Balancing the benefits of data sharing with the need for privacy and confidentiality is a complex ethical issue. Legal frameworks and ethical guidelines, such as data use agreements, data sharing agreements, and data anonymisation techniques, may be employed to ensure responsible data sharing and collaboration while protecting individuals' privacy.

Ethical considerations and legislation also focus on addressing issues of fairness and bias in health & care data. Bias in data, such as sampling bias or measurement bias, can impact the accuracy and validity of health & care findings and lead to health disparities. Ethical guidelines and legislation may require data collection methods that minimise bias and promote fairness, and also mandate transparency in reporting data and addressing potential bias in data analysis and interpretation.

Ethical considerations and legislation emphasise **transparency and accountability** in the use of health & care data. Healthcare organisations and researchers are often required to provide clear and transparent information about how data is collected, stored,

and used, and be accountable for their actions related to data use. Ethical guidelines and legislation may also require regular audits, data governance frameworks, and responsible data stewardship practices to ensure transparency and accountability in the use of health & care data.

Ethical considerations and legislation highlight the importance of **data governance and oversight** in health & care data. Data governance involves establishing policies, procedures, and frameworks to ensure responsible and ethical data use, while data oversight involves monitoring and enforcing compliance with these policies and procedures. Ethical guidelines and legislation may require healthcare organisations and researchers to establish robust data governance and oversight mechanisms to ensure responsible data practices and promote ethical data use in health & care.

Adhering to ethical principles and complying with relevant legislation in the collection, storage, use, and sharing of healthcare data is essential for ensuring responsible and ethical data practices in the health & care domain.

3.3 ECONOMY OF DATA

What we already know and what we don't yet in creation, distribution, and exchange of data and the value that is generated from it.

3.3.1 What we already know

Data is valuable: Data has become one of the most valuable resources in the world, with companies and governments seeking to collect and analyse vast amounts of it to improve their operations and decision-making.

Data creates new business opportunities: The growth of data has led to the creation of new businesses and industries that specialise in collecting, analysing, and monetising data.

Data privacy and security are important: The increased collection and exchange of data has also raised concerns about privacy and security, leading to increased regulation and the development of new technologies to protect personal information.

The role of AI: Artificial intelligence is playing a growing role in the economy of data, with machine learning algorithms used to analyse vast amounts of data to make predictions and inform decision-making.

3.3.2 What we don't yet know:

The full economic impact of data: While data has already had a significant impact on the economy, it is unclear what the full extent of its economic impact will be in the future.

The future of data privacy and regulation: The regulation of data is still evolving, and it is unclear what form it will take in the future and how it will balance the need for privacy with the benefits of data collection and analysis.

The long-term sustainability of the data economy: The rapid growth of the data economy raises questions about its long-term sustainability, including the environmental impact of data storage and processing and the ethical implications of data collection and use.

The future of work: The growth of data and AI is also changing the nature of work, leading to the automation of some jobs and the creation of new roles that require new skills. It is unclear what the future of work will look like in this new data-driven economy.

4. HUMANOME – THE INDIVIDUAL IS THE DATA CENTRE

„The Humanome is made up of two parts, 'Human' and the suffix '-ome'.
The suffix '-ome' is related to the totality of a subject and here the subject is the health of a human."

It is intended to be the most comprehensive representation of reality, and the list of data sources in the figure is not exhaustive.

„The Humanome model is a personal health profile and data repository. It is based on data that influence personal health parameters from both public and private-sector sources. The individual is placed at the centre, surrounded by a virtual assistant. Four circles represent categories, with some examples of data included, that to varying degrees affects health status. The top circle is personal behaviour, which includes data such as social media usage, physical activity, and dietary habits. The right category is biology, which covers data points like genomics, other -omics, and classical biomarkers. The bottom category is societal factors which e.g. includes data points like employment, education, and demographics. The left circle is environmental factors, covering data points such as climate, pollution, and noise. Data controls and contracts feature on the outermost circles, as all data flows in the Humanome need to comply with specific controls. They cover aspects such as interoperability, security, and safety of data, as well as transparency, traceability, and accountability for the flow and handling of data. Data contracts cover consent, donation, and sharing of data, as well as secondary use, anti-lock-in systems, and logging of data. These safety measures in data management are crucial and enable trust between individuals and institutions across sectors."

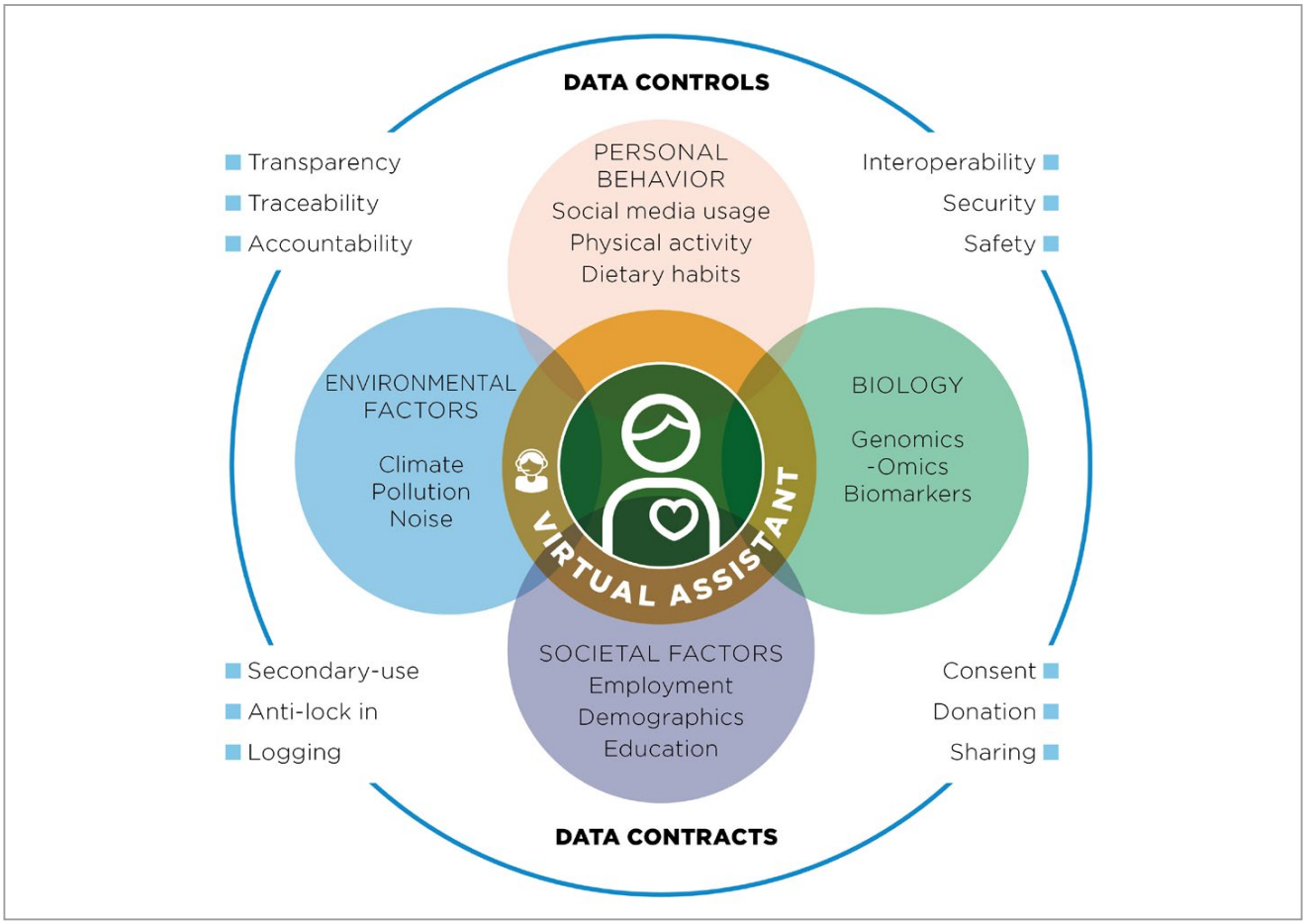


Figure 6: Individual as a data centre – the Humanome model[10]

[10] The Next Era in Global Health by Copenhagen Institute for Futures Studies, 2020.

5. ELEMENTS OF DATA DEFINITION AND ARCHITECTURE

5.1 METADATA

There are various approaches to metadata definitions and categories, We have defined five broad categories from the perspective of data utilisation and the role of these categories in accurately positioning of a data point within the complete data space.

I. Provenance https://www.w3.org/TR/prov-overview/ with time stamp

II. Content, i.e. code systems, format

III. Access Route, database or algorithm and its security

IV. Context, e.g. ontology/knowledge graph

V. Data connections in the data space, e.g. relationships and transfer modes

Metadata catalogue under development for secondary use in health by the HealthData@EU Pilot project (https://ehds2pilot.eu):
DCAT-AP https://www.w3.org/TR/vocab-dcat/

5.2 DATA QUALITY

Data quality is a pivotal facet in the realm of data science and projections, wielding a profound impact on the soundness and dependability of any data-driven analysis. The accuracy, reliability, and integrity of data stand as pivotal determinants, directly shaping the validity and robustness of analytical outcomes.

Accurate data forms the bedrock of reliable projections and data science analyses. It hinges on the fidelity of data in representing the true value or state of the phenomenon under scrutiny. Inaccuracy in data can engender flawed projections and misleading conclusions, exemplified by the potential pitfalls of inaccurate data in forecasting future trends through time series analysis.

The significance of complete data emerges as a critical factor for comprehensive analysis and precise projections. Completeness denotes the presence of all requisite data points or variables essential for a thorough analysis. The absence or incompleteness of data can introduce biases into projections, necessitating the use of data imputation techniques. However, the reliability of imputed data becomes a concern if not managed meticulously.

Data consistency assumes a paramount role in ensuring the reliability and validity of projections. It revolves around the uniformity and coherence of data across diverse sources, variables, or time periods. Inconsistent data can yield conflicting results or inaccurate projections, underscoring the importance of employing data validation and verification techniques to rectify disparities.

Timeliness in data is indispensable for crafting up-to-date projections and conducting accurate analyses. The freshness and relevance of data to the ongoing analysis define data timeliness. Stale or outdated data can result in obsolete projections or inaccurate

insights, necessitating vigilant management of data collection and processing timelines.

The crux of meaningful projections and reliable analysis lies in the use of relevant and reliable data. Relevance pertains to the applicability of data to the analysis at hand, while reliability hinges on the trustworthiness and credibility of data, encompassing its source, methodology, and quality assurance measures. Utilising irrelevant or unreliable data risks the creation of inaccurate projections and misleading conclusions.

Addressing data bias and ensuring fairness are paramount in the ethical landscape of data science and projections. Data bias denotes systematic errors or distortions in data that can skew projections or foster discriminatory outcomes. Fairness, on the other hand, revolves around the equitable treatment of different groups or individuals in data-driven analysis. Failure to mitigate bias or ensure fairness can lead to ethically fraught decision-making and legal repercussions.

Data security and privacy represent linchpins in the tapestry of data quality for projections and data science. Safeguarding data during storage, processing, and sharing, along with upholding individuals' privacy rights, stands as a critical imperative. Breaches or violations in this realm can sow the seeds of data quality issues and erode trust in projections or data-driven analyses.

Additionally, the institutional environment surrounding data—comprising information on origin, governance arrangements, and the institutions involved in data collection, processing, and retention—plays a pivotal role in determining data quality. It adds another layer of contextual understanding to the data's reliability and relevance.

Data accessibility, denoting the ease with which data can be accessed and the means and locations for such access, further contributes to the overarching narrative of data quality. Finally, data interoperability, or the provision of information that facilitates the proper understanding and utilisation of data, acts as a bridge to ensure that data can seamlessly integrate into diverse analytical frameworks.

Current practice is characterised by numerous biases.[11]
Results: ..."We identified 13 possible sources of bias. Four of them are related to the organization of a healthcare system, whereas some are of a more technical nature."
Conclusions: "There are a substantial number of possible sources of bias; very little is known about the size and direction of their impact. However, anyone that uses or reuses data that were recorded as part of the healthcare process (such as researchers and clinicians) should be aware of the associated data collection process and environmental influences that can affect the quality of the data."

Data quality framework project increases use of data: https://www.stat.fi/org/tilastokeskus/vuosiohjelma_en/data-quality-framework-project-increases-use-of-data.html

A good practice is to combine DCAP-AP 2.0 with PROV-O (provenance) to add both dimensions to data quality and lineage (how, where, when, who is responsible for the data-set published) - https://www.w3.org/TR/prov-o/ and https://www.w3.org/TR/vocab-dqv/

[11] Verheij, R.A., Curcin, V., Delaney, B.C., McGilchrist, M.M. Possible sources of bias in primary care electronic health record data use and reuse. Journal of Medical Internet Research: 2018, 20(5), e185, https://pubmed.ncbi.nlm.nih.gov/29844010/

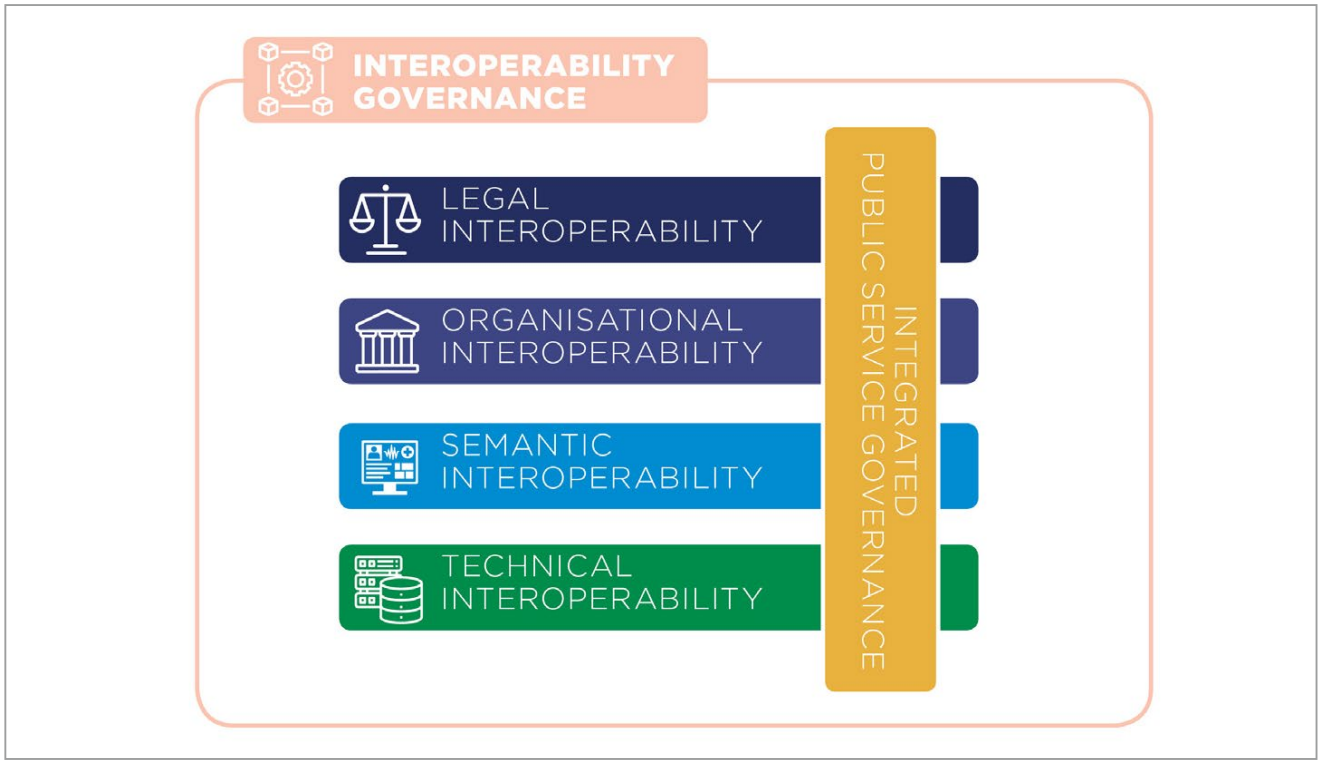


Figure 7: Interoperability model of the European Interoperability Framework[12]

Existing solutions:
National and international standards for semantic interoperability:

- monolingual and multilingual terminologies (SNOMED CT, UCUM, EDQM)
- international coding systems (ICD, LOINC, ATC)
- information models (HL7, CDA, FHIR, OMOP-CDM)
- data object (JSON-LD structure and a GraphQL end-point)

Supporting technology for technical interoperability:

- terminology servers: increasingly used
- Extract-Transform-Load (ETL) tools: graphical UI for file conversions, schema mappings, data transformations, sometimes code mappings: widely used;
- information extraction tools: never in care, somewhat in research, more often for accounting.

[12] https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/3-interoperability-layers

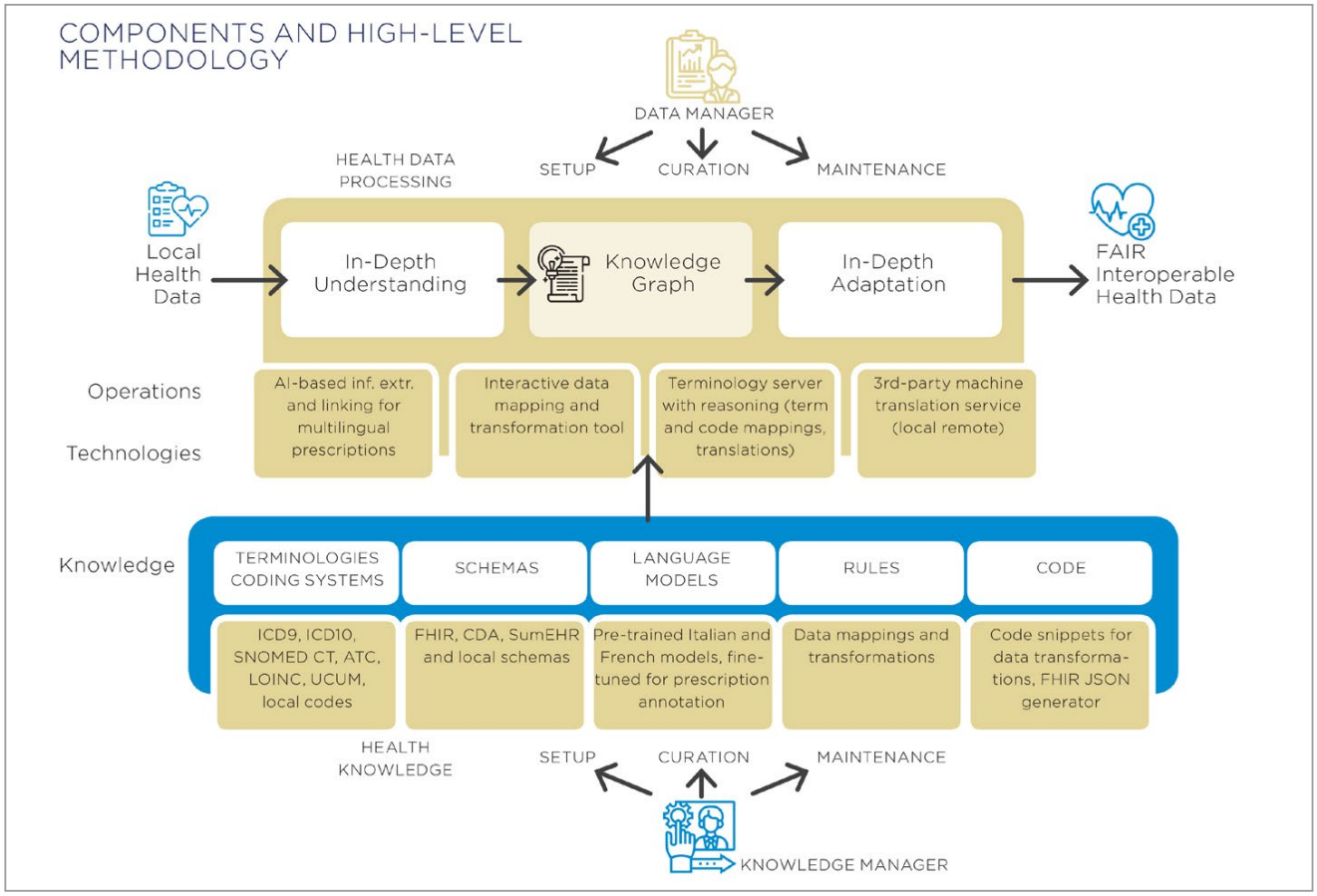


Figure 8: Mapping methodology for semantic interoperability[13]

There is a growing need for mapping data from new sources, especially as more user-provided data is generated.
User generated data
Smart things (health & wellbeing devices and services) IoT and use of blockchain (contracts) to coordinate medication, treatment and personal health data.
https://github.com/rpi-scales/BlockIoT
https://solidproject.org/apps
Patient Reported Outcome Measures (PROM) and Patient Reported Experience Measures (PREM) where there are a lot to learn from behavioural sciences, also observational and questioning techniques need to be improved.

Organisational interoperability, legal interoperability and governance vary widely from country to country.

5.4 KNOWLEDGE GRAPH

A knowledge graph in health and care is a type of database that uses a semantic structure to organise and connect different types of health-related information. It consists of nodes (entities) that represent concepts, such as diseases, symptoms, treatments, and medications, and edges (relationships) that represent the connections between these concepts. This allows the knowledge graph to represent complex relationships between

[13]InteropEHRate project. Simone Bocca, Gábor Bella, Yamini Chandrashekar, 2022.

different types of health information in a structured and meaningful way.

The knowledge graph can be used to support a range of healthcare applications, such as clinical decision support, personalised medicine, and population health management. For example, it can be used to identify patterns and trends in patient data that may indicate a risk for certain diseases or conditions, or to suggest appropriate treatment options based on a patient's specific health profile.

Examples of current practice:
Zero shot slot filling with Dense Passage Retrieval (DPR) and Retrieval Augmented Generation (RAG)[14] explores the use of language models and knowledge graphs to improve the accuracy of natural language in conversational agents. The paper specifically focuses on the use of Dense Passage Retrieval (DPR) and Retrieval Augmented Generation (RAG) techniques for zero-shot slot filling, a task in which the conversational agent[15/16] must understand and fill in missing information in a user's query.

These studies highlights the potential of combining large-scale language models, such as BERT and GPT-3, with knowledge modelling to improve the accuracy and efficiency of zero-shot slot filling. By leveraging the knowledge contained in the knowledge graph, the conversational agent is able to better understand the context of the user's query and provide more accurate and relevant responses.

Recent developments emphasise the importance of including human-in-the-loop thinking in the development of conversational agents, as this can help to ensure that the agent is providing useful and relevant information to the user. Overall, the research presented in these papers has the potential to enhance the accuracy and efficiency of natural language understanding in conversational agents, ultimately resulting in a better user experience.

"The second workshop in the Knowledge Graph Conference"[17] focused on Personal Health Knowledge Graphs, which are knowledge graphs that are specifically designed to capture and represent an individual's health-related data and information. The workshop featured presentations and discussions on various topics related to personal health knowledge graphs, including data modelling, data integration, privacy and security, and applications in clinical decision-making.
Speakers highlighted the potential benefits of personal health knowledge graphs, such as improved patient outcomes, more personalised and effective care, and increased patient engagement and empowerment. However, they also noted the challenges associated with developing and implementing these knowledge graphs, such as the need

[14]Glass, Michael & Rossiello, Gaetano & Gliozzo, Alfio. (2021): Zero-shot Slot Filling with DPR and RAG. https://doi.org/10.48550/arXiv.2104.08610
[15]A. Meloni, S. Angioni, A. Salatino, F. Osborne, D. Reforgiato Recupero and E. Motta, "Integrating Conversational Agents and Knowledge Graphs Within the Scholarly Domain," in IEEE Access, vol. 11, pp. 22468-22489, 2023, doi: 10.1109/ACCESS.2023.3253388.
[16]Demetriadis, S., Dimitriadis, Y. (2023). Conversational Agents and Language Models that Learn from Human Dialogues to Support Design Thinking. In: Frasson, C., Mylonas, P., Troussas, C. (eds) Augmented Intelligence and Intelligent Tutoring Systems. ITS 2023. Lecture Notes in Computer Science, vol 13891. Springer, Cham. https://doi.org/10.1007/978-3-031-32883-1_60
[17]DThe Personal Health Knowledge Graph Workshop.4 May 2021, Virtual. https://phkg.github.io/

for standardised data models and interoperability, data quality and completeness, and addressing privacy and security concerns.

After defining the concept of the Humanome in Chapter 4, which seeks to encapsulate the entirety of human existence and aims to be the most comprehensive data-driven representation of reality, we introduce the term 'Humanome Knowledge Graphs.' These graphs are designed to serve as knowledge models within a multi-dimensional data space, specifically tailored to represent various life situations and conditions. In essence, these knowledge graphs are intricate and dynamic data structures that provide a rich and multifaceted understanding of human experiences and states, enabling a more detailed exploration of the diverse aspects of human life within a comprehensive data framework.

5.5 DATA INTEGRATION

F.A.I.R-fuelled Reference Data Integration & look-up service[18], which is a data integration platform that utilises the F.A.I.R. principles (Findable, Accessible, Interoperable, and Reusable) to enable the sharing and integration of reference data across different systems and organisations. This platform provides a central repository for reference data, such as codes, classifications, and vocabularies, that are commonly used in healthcare and other domains.

The look-up service component of the platform ensures the integrity and security of the reference data by preventing unauthorised changes or deletions. This is particularly important in healthcare, where accurate and consistent use of reference data is critical for ensuring patient safety and effective communication between providers.

6. LIMITATIONS, HINDRANCES

6.1 INCOMPLETE DATA FLOW

Current healthcare representation of the data loop (see 3. Mechanism – data loop) is the trio of 'research zone', 'database zone' and 'care zone' according to Robert Verheij et. al.[19]. The different steps are involved in the reuse of routinely recorded health data in research, but also in every day healthcare practice. Each of these seven steps are subject to possible distortion of the image of reality to be seen by the end user of this data. Each of these steps require adequate metadata in order to allow the end users to understand what the data is actually representing. Absence of metadata will increase the risk of bias in the outcomes of the data. Information, knowledge and wisdom generated by any data of which the fitness for purpose is not known is due to be false.

That is why we have to invest in metadata and in knowledge about data at the end users.

[18] https://vimeo.com/529328116/732160e4b4

[19] Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. J Med Internet Res. 2018 May 29;20(5):e185. doi: 10.2196/jmir.9134https://doi.org/10.48550/arXiv.2104.08610

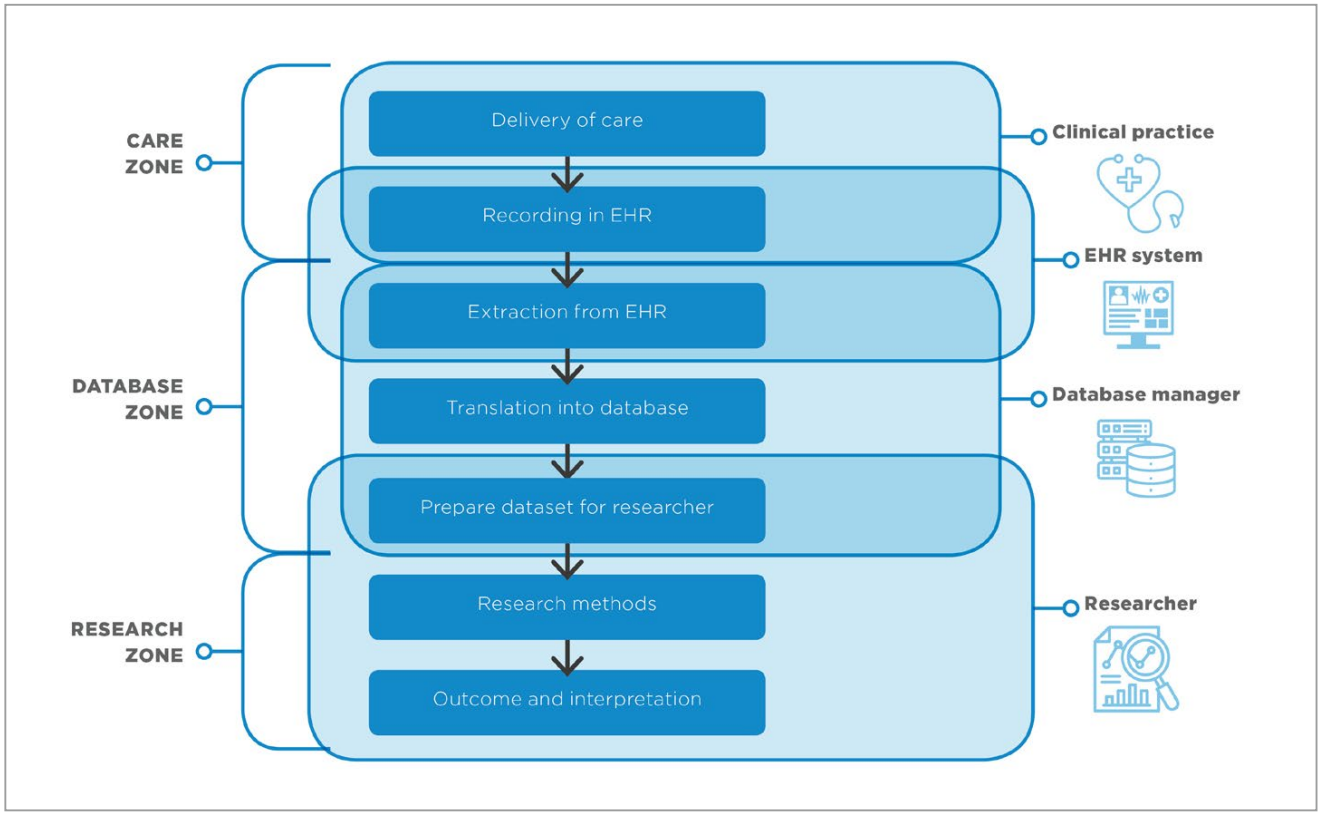


Figure 9: Steps and actors involved in the data flow between the delivery of care and applications reusing the data. EHR: electronic health record

Factors that might affect the fitness for purpose and we still know little about and for which we yet lack adequate methodologies to measure them:

- Healthcare system bias, emanating from:
- Reimbursement system, pay for performance parameters
- Role of healthcare provider in the healthcare system;
- Professional clinical guidelines
- Difficulties of access by patients to their records
- Data sharing between healthcare providers
- Legal and ethical barriers for health data reuse.
- Practice workload, administrative burden
- Variations between electronic health record (EHR) system functionalities and lay-out
- Incomplete coding systems and thesauruses
- Poor knowledge and education regarding the use of EHR systems
- Inadequate data extraction tools
- Deficiencies in data processing—re-databasing
- Inadequate research dataset preparation
- Inadequate research methodologies

The figure below also refers to a 'care zone', a 'database zone', and a 'research zone' and demonstrates data transitions between health professional's domain into a database environment, and into a research environment.

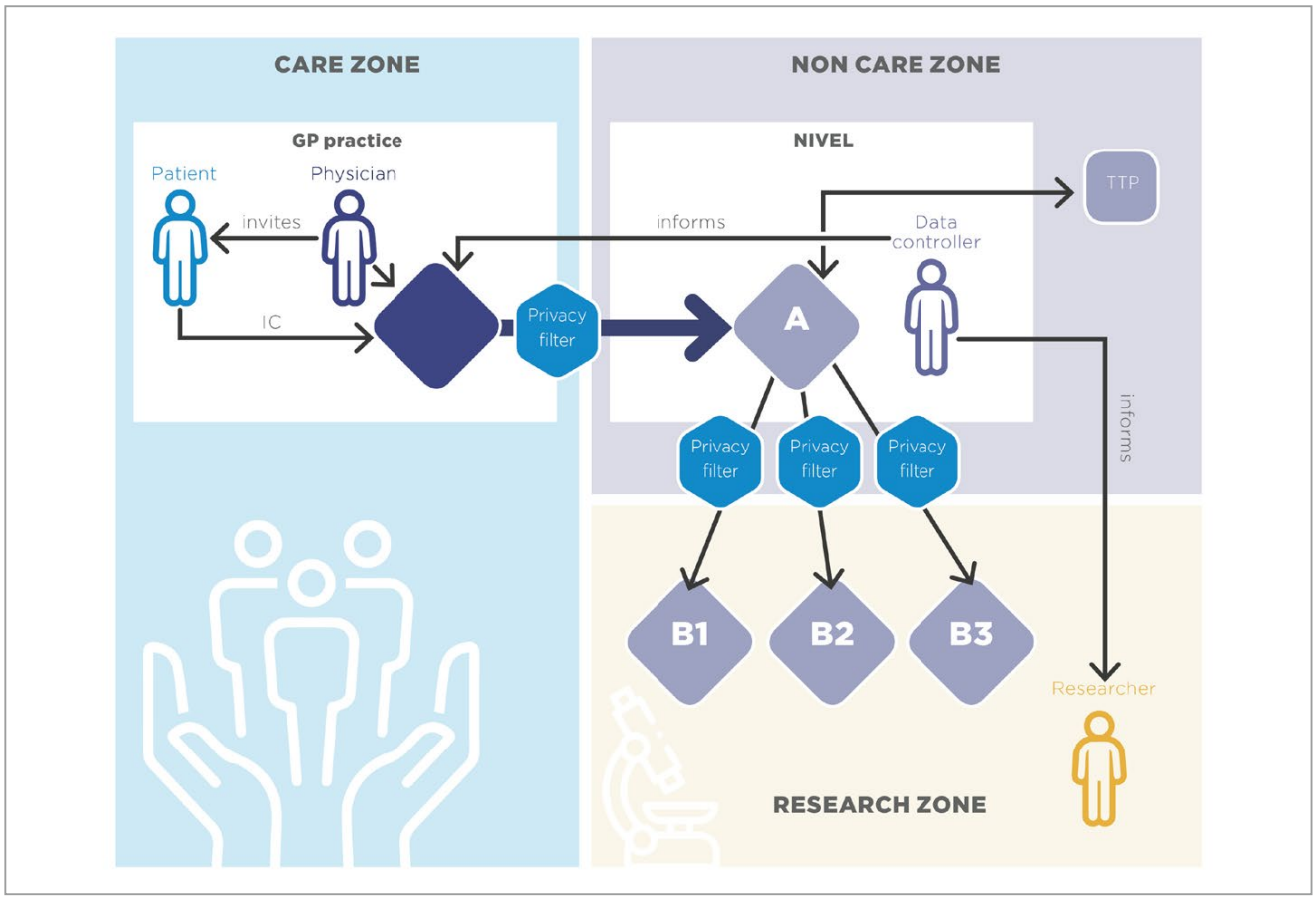


Figure 10: Representation of the example depicting data exchange and research done with NIVEL-PCD. Regularly patient data is transferred from the GP to the NIVEL database, where data is stored with pseudonyms (database A). Researchers can obtain additionally pseudonymised and aggregated extracted data sets (database B1-B3); only a subset of patient data that is doubly pseudonymised (practically anonymised) can be analysed by the researcher; (IC = informed consent, TTP = Trusted Third Party)[20].

6.2 OUTLIERS

Outliers in data science significantly deviate from the normal pattern of the rest of the data and can have a significant impact on data analysis and modelling. Outliers can distort projections and lead to inaccurate predictions, introducing bias in data analysis, and impacting the interpretability and explainability of data science models. Additionally, outliers can pose challenges in data quality and data cleaning, affecting the accuracy and reliability of data. Models that are sensitive to outliers may not generalise well to new data or real-world scenarios, impacting decision-making and risk management based on data-driven insights. Properly addressing outliers through careful data cleaning, model robustness testing, and interpretability can help ensure the accuracy, reliability,

[20]Kuchinke W, Ohmann C, Verheij RA, van Veen EB, Arvanitis TN, Taweel A, Delaney BC. A standardised graphic method for describing data privacy frameworks in primary care research using a flexible zone model. Int J Med Inform. 2014 Dec;83(12):941-57. doi: 10.1016/j.ijmedinf.2014.08.009. Epub 2014 Sep 3.

and interpretability of data science models and projections. Domain expertise and sound data practices are essential in dealing with outliers and their impact on data-driven insights.

6.3 INCORRECT DATA

Incorrect data can lead to inaccurate projections, flawed data analysis, and invalid conclusions in data science. Such mistakes undermine the trustworthiness and reliability of projections and analysis, and can result in costly errors, reputational risks, and ethical concerns. Incorrect data in projections can lead to misleading forecasts, while flawed analysis can lead to biased predictions or classifications. Invalid conclusions can misrepresent reality and lead to incorrect insights or recommendations, potentially resulting in financial losses, reputational risks, and ethical concerns. To mitigate the issues associated with incorrect data in projections and data science, it is essential to implement robust data quality management practices, including data validation, verification, and cleansing. Regular data quality checks, data cleansing processes, and validation techniques can help identify and rectify incorrect data before it is used for analysis or projections. Ensuring data accuracy, consistency, and reliability is critical for generating trustworthy and meaningful insights from data-driven analysis.

6.4 SYSTEMATIC ERRORS

Systematic errors, also called bias, can have significant impacts on projections and data analysis, as biased projections, with which over- or under-reporting of data can result in inaccurate future trend estimates. Systematic errors can distort data analysis and result in incorrect or unfair outcomes, biased data can lead to flawed decision-making in various domains, stakeholders may lose confidence in the results due to skewed data, leading to scepticism. Biased data can lead to unequal treatment or biased outcomes, and in regulated domains can result in non-compliance and legal disputes.

To mitigate the issues associated with systematic errors in projections and data science, it is essential to implement rigorous data validation, verification, and cleansing processes. Regular data quality checks, identification, and correction of biases in data, and thorough validation of analysis results can help identify and address systematic errors. Ensuring data accuracy, integrity, and fairness are critical for generating reliable and unbiased insights from data-driven analysis. Additionally, employing diverse and representative data sources, using robust statistical techniques, and promoting transparency in data collection and analysis can also help mitigate the impact of systematic errors in projections and data science.

6.5 SMALL DATASETS

Small stand-alone datasets pose challenges for projections and data science analysis. These include limited statistical power, higher variability, increased risk of overfitting, sampling bias, limited scope and generalisability, and reduced confidence and reliability. These issues can result in less accurate projections and models, unreliable and inconsistent results, and limited applicability of findings. Stakeholders may question the validity

of the results, leading to lower confidence and hindering effective decision-making and planning.

To mitigate the issues associated with smaller datasets in projections and data science, it is important to be mindful of the limitations of small sample sizes and take appropriate measures. This can include using statistical techniques that are suitable for small datasets, such as bootstrapping or resampling, validating results through robust cross-validation, and being cautious in making extrapolations or generalisations. Ensuring data quality, addressing sampling bias, and carefully interpreting and validating the findings can help enhance the reliability and accuracy of projections or data science analysis based on smaller datasets. Additionally, considering complementary data sources, employing domain expertise, and leveraging expert judgment can provide additional insights and mitigate the limitations of smaller datasets in projections and data science.

7. GOOD PRACTICES

Effective healthcare is built on a foundation of good practices that have been tried and tested over time. In this chapter, we will explore best practices that have demonstrated the effective use of data in different aspects of practice.

7.1 METADATA REGISTRY

The Australian Government Online Metadata Registry (METEOR)[21]

Australia's health system is complex, and so are the structures, processes and standards that support it. Reflecting the federated governance structures of Australia, the overarching regulation and funding of Australia's health system is the joint responsibility of the Australian Government (Federal Government) and the State and Territory Governments (Jurisdictions).

Central to Australia's health system is Medicare – a universal public health insurance scheme. Medicare covers the full cost of services through public hospitals as well as providing rebates for the costs of some medical services and prescription pharmaceuticals. Medicare is funded by the Federal Government through a means-tested tax. Services not covered by Medicare, or those provided to private patients through hospitals, general and specialist services, or pharmacies are paid for either by full or partial patient contribution. Patients with private health insurance are charged the difference between their coverage and the full fee. The amount of difference can vary significantly depending on the type of insurance held and the service and provider used.

Provision of health services in Australia is shared across both the government and private sectors, with governments largely responsible for public services and the private sector (both for-profit and not-for-profit) for private services. In addition, some health services are delivered by local community services and out-reach clinics funded through specialist health programs. This diversity in service providers contributes to a wide mix of health information and data collection methodologies. Currently, there is no standardised infrastructure for health data across Australia.

[21] https://www.aihw.gov.au/about-our-data/accessing-data-through-the-aihw/metadata-standards

The Australian Institute of Health and Welfare (AIHW) and METEOR

Much of Australia's health data is administrative, designed to support operational needs of Australia's health system and measure activity-based funding. These administrative datasets are usually collected at the person-level (unit record based) by service providers and held by the relevant jurisdictional government department. As part of the health system funding agreements across all Australian Governments, a sub-set of jurisdictional health data is provided to the Federal Government – either directly to the Australian Institute of Health and Welfare (AIHW) or through the Commonwealth Department of Health and Aged Care (DoHAC). Jurisdictional data are combined into a national collection that supports performance monitoring of Australia's health system.

Administrative data in each jurisdiction are created to support local practices, using local data specifications – without agreed, consistent, national metadata for reporting purposes, they are unsuitable for comparison or analysis. Metadata describes how data are defined, collected, and structured. They provide meaning and context to data. Robust metadata help interpret data and supports national consistency and cross-collection comparisons, while ensuring data users and data custodians share a common understanding of data. The AIHW provides the mechanisms and infrastructure for the development of national data standards. Data standards are metadata that have been through a formal review and quality assurance process and are endorsed by an authoritative body.

The AIHW works in collaboration with jurisdictions and other agencies who support the health system – for example, the Independent Hospital and Aged Care Pricing Authority (IHACPA – to develop, review and endorse data standards specific to national collections. These are called Data Set Specifications (DSSs). Australian health DSSs come in three main types:

1. National minimum data set (NMDS) – a minimum set of data elements agreed for mandatory collection and reporting at a national level.
2. National best endeavours data set (NBEDS) – a set of data elements that is not mandated for collection, but where there is a commitment to provide data nationally on a best endeavours basis.
3. National best practice data set (NBPDS) – a set of data elements that is not mandated for collection but is recommended as best practice.

NMDSs include sets of data elements that have been agreed to by all jurisdictions nationally. An NBEDS generally works as a 'road map' towards an NMDS – the jurisdictions agree to the importance of the collection as a whole, but (often due to ICT limitations) may not yet be able to collect or extract the data consistently. NBPDSs provide guidance on the preferred way of collecting data variables where no agreed national data collection exists. Note that these agreed national data standards do not preclude the jurisdictions from collecting additional, or more granular, data for their own purposes.

DSSs support the national indicators used to monitor and performance measure Australia's health system. The national indicators provide the specifications for statistical

measures that perform calculations on the raw data from AIHW's data collections. To that end, they link to the data elements within DSSs that describe this data, allowing users to understand more clearly how calculations are made. National indicators collected over time allow Australian policymakers to assess the extent to which outcomes are being achieved. All nationally endorsed health DSSs are publicly available on AIHW's METEOR – the Australian Government Metadata Online Registry for health and welfare data.

METEOR and preceding versions have been an essential part of Australia's health information infrastructure for more than 20 years. The metamodel for describing data standards in METEOR is based on the International Organization for Standardization and the International Electrotechnical Commission 11179 Information technology — Metadata registries (ISO/IEC 11179). The metamodel describes variables using data element structures.

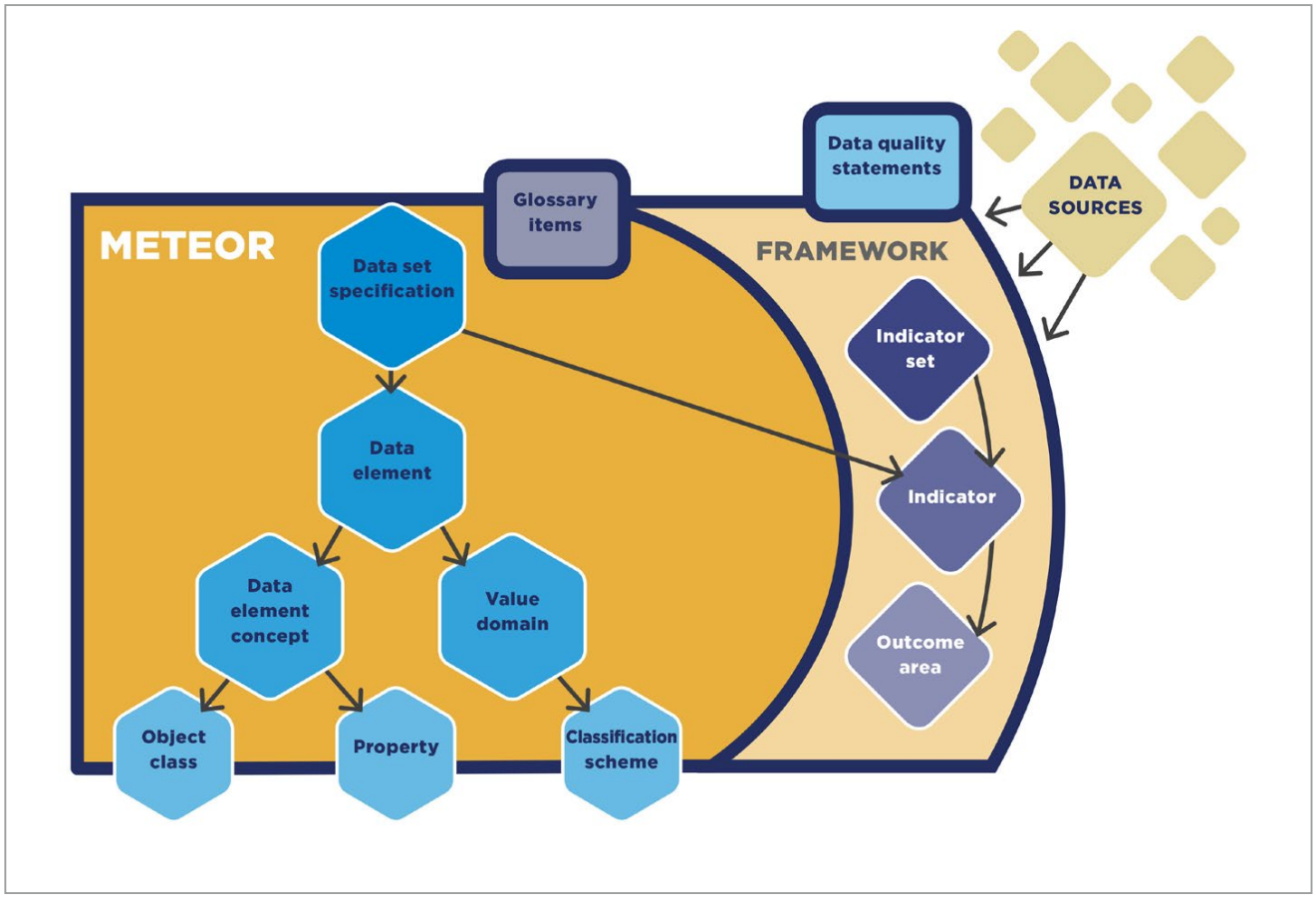


Figure 11: METEOR metamodel and relationships[22]

ISO/IEC 11179 caters for metadata extensions that meet changing user expectations and the unique needs of the Australian health data environment. AIHW has exercised its significant metadata expertise to engineer extensions such as indicator sets, data sources, and data quality statements. AIHW pioneered the development of DSS metadata which was subsequently recognised and incorporated in the ISO/IEC 11179 standards suite.

[22] https://meteor.aihw.gov.au/content/268284

Underpinned by strong metadata governance and quality assurance frameworks, METEOR supports the data value chain from initial collection through to submission, validation, analysis, interpretation, and reporting. METEOR includes both prescriptive metadata (i.e. definitions of data that are to be collected in an upcoming reporting period) as well descriptive metadata (data that have been collected in the past.) Through accessible, consistent, and comparable national data, METEOR supports the FAIR data principles – findable, accessible, interoperable, and reusable data, making it easier for organisations to build quality evidence-based policies, frameworks, and reporting mechanisms. Metadata in METEOR is publicly available (findable/accessible) and relies on the principle of 'define once, use often' (interoperable/reusable.)

Supporting data comparability, integration and quality

National standards ensure information that is collected by jurisdictions are consistent, accurate, and useful. For example, if one collection defines 'young adults' as people aged between 20 and 25, but another defines them as people aged between 18 and 22, the 'young adult' data can't be compared because the age groups across the collections differ. Accessible robust standards provide users with this essential context for interpretation. National standardised metadata, like DSSs, can also drive improvements to collections, by providing a structured mechanism for metadata development – enabling utilisation of data collections as part of a broader asset through linkage. AIHW is a nationally accredited Data Integration Authority, the Data Linkage Unit (DLU) at AIHW, use data standards to support national linkage projects. For example, linkage of person-level hospital data with Medicare data, and cause of death data, enabling analysis of pathway patterns through Australia's hospital and health systems.

A key value of METEOR is that it provides clarity on what the current data means, as well as a record of changes to the data over time. Understanding the provenance of data provides users with the opportunity to compare discrete data sets across services, geographic areas, and sectors. METEOR also includes Data Quality Statements (DQSs), these are written in a standard format and provide information about aspects of the data that has been collected, such as institutional environment, timeliness, accessibility, interpretability, relevance, accuracy, and coherence. A DQS helps users to understand important data limitations and make informed judgements when they use the data collection described. AIHW policy requires DQSs for collections where AIHW is the data custodian. DQSs go through an expert review process and are available to the public on METEOR.

To ensure all the metadata in METEOR – including DSSs, Indicators, and DQSs are high quality and fit-for-purpose, AIHW requires a process of review and endorsement for all data standards published in METEOR. Registration Authorities (RAs) are the governance bodies responsible for each sector's data standards – RAs review recommendations made by relevant data or information committees and provide formal approval for assigning the 'Standard' status to a metadata item. In Australia's health sector, the RA is the AIHW, and the expert committee is the National Health Data and Information Stan-

dards Committee (NHDISC). NHDISC has representatives from all the jurisdictions, in addition to a range of key stakeholders from the Australian Government. NHDISC has three annual cycles, during which new metadata and updates to existing metadata are reviewed and endorsed.

METEOR and the information and data standards processes supporting it are essential to the quality improvement of Australia's health information, as a result, work in this space remains ongoing. The AIHW are currently working on a range of system improvements to METEOR to enable improved secure automation of extracts and inputs through the API development program, along with exploring integration with AIHW's data validation tool. METEOR is the authoritative source of truth for the valid values of data items collected at AIHW. The metadata in METEOR defines permissible values, data type, character length, and format (among other information about data). This information is defined and published on METEOR prospectively; before data are collected or submitted. The AIHW is exploring an opportunity to harvest efficiency gains by leveraging the valid values in METEOR to perform initial checks on incoming data submissions. This prevents duplication of effort in re-defining and applying valid data values using a separate data validation system. Metadata management capability underpins data validation and is fundamental to the entire data value chain from specification to submission, analysis and reporting. As Australia's uptake of the person-centred My Health Record increases, AIHW is reviewing options for how METEOR can best support this vital asset. These type of continuous development and enhancement of the METEOR platform and processes that support it, demonstrate AIHW's commitment to advancing Australia's digital health maturity and best-practice information standards.

7.2 IMPROVED DATA UTILISATION

Data Mesh Principles and Logical Architecture[23]

Our aspiration to augment and improve every aspect of business and life with data, demands a paradigm shift in how we manage data at scale. While the technology advances of the past decade have addressed the scale of volume of data and data processing compute, they have failed to address scale in other dimensions: changes in the data landscape, proliferation of sources of data, diversity of data use cases and users, and speed of response to change. Data mesh addresses these dimensions, founded in four principles: domain-oriented decentralised data ownership and architecture, data as a product, self-serve data infrastructure as a platform, and federated computational governance. Each principle drives a new logical view of the technical architecture and organisational structure.

Data mesh should also be linked with a semantic mesh, which provides a standard framework for the knowledge graphs.

[23] https://martinfowler.com/articles/data-mesh-principles.html

7.3 THE INDIVIDUAL AS A DATA CENTRE

(see also 4. Humanome above)

7.3.1 MYDATA[24]

The MyData project is an initiative that promotes the idea of individuals being in control of their personal data, including health data, and using it for their own benefit. From the perspective of the individual as a health data centre, the MyData project emphasises empowering individuals to be the custodians of their own health data and to share it with trusted entities for specific purposes.

In the context of health data, the MyData project envisions individuals having access to their own health data from various sources, such as electronic health records (EHRs), wearable devices, self-reported data, and genetic data, among others. Individuals can store and manage their health data in a personal data store, which is a secure online platform or tool that allows them to collect, store, and control access to their data.
The MyData project promotes the idea of individuals being able to share their health data with different stakeholders, such as healthcare providers, researchers, and other entities, based on their informed consent and preferences. Individuals can define the purposes for which their health data can be used, specify the duration of data access, and revoke access at any time. This puts individuals in control of their health data and enables them to decide who has access to their data and for what purposes.

The MyData project also emphasises the importance of data interoperability and standardisation, allowing individuals to aggregate and share their health data across different platforms and systems in a standardised format. This enables individuals to have a holistic view of their health data and promotes seamless data exchange between different stakeholders involved in their healthcare.

The MyData project also advocates for transparency, accountability, and data ethics in health data management. It encourages organisations that collect and process health data to be transparent about their data practices, respect individuals' privacy, and adhere to ethical principles in data management, such as data minimisation, purpose limitation, and data security.

Overall, the MyData project promotes the concept of individuals as empowered health data centres, where individuals have control over their health data, can share it for specific purposes based on their preferences, and are empowered to make informed decisions about their health and well-being. It aims to foster a more patient-centric approach to health data management, where individuals are active participants in the process and have the ability to harness the potential of their health data for their own benefit.

7.3.2 DATAVAULTS[25]

DataVaults aims to deliver a novel framework and architecture that leverages personal data, coming from diverse sources (sensors, IoT, wearables, data APIs, historical data,

[24] https://www.mydata.org
[25] https://www.datavaults.eu

social network data, activity trackers, health records, demographic profiles, etc.) to help individuals construct their unified personal data hub, collect at a single point all of their personal data in a secure and trusted manner, and retain ownership and control on what to share and with whom, receiving also compensation for the artefacts they place at the disposal of other third parties. In turn, third party organisations (companies, public sector, NGOs, etc.) arrive at a position where they can request and get access to tons of personal data, which can complement the ones they already manage and that can be used for generating more efficient, effective and value added services, engaging with individuals into an entirely new way for data sharing, which generates trust and an increased feeling of collaboration, as the data owners (e.g. individuals) become the centre of attention and the most important partner and collaborator of those third parties.

At the core of DataVaults lies the DataVaults personal data value chain which could from now on be seen as a multi-sided and multi-tier ecosystem governed and regulated by smart contracts to safeguard personal data ownership, privacy and usage and attribute value to all entities that generate value within this chain and especially data owners. DataVaults will deliver a framework and platform that will set, sustain and mobilise this ever-growing ecosystem for personal data sharing and for enhanced collaboration between those who own data (data owners) and those who seek data (data seekers).

7.4 REGULATION

Experiences on the Finnish regulation on national metadata descriptions of health data

Background

The Finnish legislation for secondary use of health and social data entered into force and the Finnish Social and Health Data Permit Authority Findata started operations in 2019. In February 2021 Findata issued The Regulation of the Health and Social Data Authority: Data content, concepts and data structured for data descriptions with aim to steer the metadata description work on the data under the secondary use legislation.

Data descriptions

All data holders are responsible for describing their own data in the common metadata catalogue for health data, Aineistokatalogi / Data resources catalogue, in a manner specified by Findata's regulation. The metadata descriptions are for now made manually in a metadata description editor tool Aineistoeditori / Data resource editor. CSV import is possible for importing variable descriptions. The tools are maintained by Findata and Finnish Institute for Health and Welfare (THL).

The Regulation on data descriptions requires data to be described in data resource (study), dataset, variable and code list levels. Also, variable level information on possible considerations related to data quality is required. It is possible to make the data descriptions in three languages, Finnish, Swedish and English, although only the primary language (usually Finnish) is required. By the autumn of 2023 many data holders have already described their data. Currently there are descriptions by approximately 40 data holders.

Good practices

Starting the national metadata description work and the deployment of the systems were made easier by the groundwork done in the Isaacus project (years 2016 – 2018) developing the metadata model and the metadata description tools and testing them by describing different types of data. The metadata model was designed to support the secondary use of data and it was mapped to some key metadata standards like GSIM and DDI-L to support the interoperability. It was also considered beneficial that the metadata description systems were already in production use in THL, THL had done some further development in them, and that several data holders had piloted making their data descriptions before the start of Findata.

The data holders have been instructed to start their metadata work by describing the data that is the most requested from them for secondary use and after that expand the work to more data resources and more detailed descriptions. Some of the data holders didn't have any initial descriptions to start from and some had already described their data using their own systems. The data holders are also encouraged to do the metadata work because it will make their own work easier and save time from both them and Findata in giving consultation service to the customers. Findata and THL support the data holders in their metadata description work by providing free courses on using the metadata description tools, as well as by giving them detailed written instructions.

Challenges

It is important to understand that good and detailed variable level metadata descriptions require time and effort. The work is human resource intensive and making good descriptions, especially including the information on data quality, requires different types of knowledge of the data. Some of the healthcare providers have commented that they don't actually have the detailed descriptions of the data they own because their patient and customer information system providers own their system's data models. In these cases, the healthcare providers have bought the metadata description work from the ICT system providers.

Making good metadata descriptions also requires good and functioning tools and systems. The metadata tools in use in Finland are being developed further but there have still been some technical problems. All of the planned functions are not yet in use, like the possibility to import descriptions via API.

The consistency in the descriptions is also a challenge. For example, finding a clear and uniform way to define what is a data resource and what defines the data sets attached to it isn't easy. The approach has been to encourage the data holders to describe their data first and to start further centralised quality controls possibly later on. Nevertheless, it has been found to be a good approach to first get a vast number of descriptions and later, when more experience has been gained, to start harmonising them.

7.5 QUALITY PATIENT REGISTRY

Hungarian Haematology Patient Registry with the OMOP Common Data Model, the SNOMED CT - ICD-10 mapping and the JSON data object approach.

ClinRegSys dynamic register engine

The ClinRegSys register engine is a dynamic, run-time reconfigurable general health data acquisition register engine developed by eHealth Software Solutions. ClinRegSys contains three basic entities: the Register, the Register Schema, and the Register Entry.

The Registry is a high-level entity describing the health registry, contains the basic data collection information and summarises the schemas.

The Registry schema is a dynamically created entity that describes the data content of the registry in the form of data structures. A register can have more than one register schema, the schemas are distinguished by their version. The schema contains the format of the register entries and the validation requirements. The schema may contain dynamic query elements. Most often, a schema will also include a fillable form to assist data sources who cannot upload data to the register through any other electronic channel. Schemas in the ClinRegSys system must be provided in the JSON format RFC8259. The Register entry stores the data added to the register in the format – structured – defined by the schemas.

The Hungarian 49/2018 decree of Ministry of Human Resources states the registers, the professional owner institutes of each register, the common format, and the basic content of the register entries in general. The pre-defined data content of each register entry is:

- Institute data
- Patient data
- Case data (include ICD-10, WHO Code, mCODE)
- Disease specific data (based of certain ICD-10 groups)
- Outcome data

After this general decree each register owner creates a methodology letter about the detailed, profession-specific, or disease-specific data content of their registries.

For example, the data model of the National Registry of Haematology Diseases contains 230 different unique data fields for five ICD-10 groups (ML, CLL, MM, HL, NHL). From these 49 fields are fix data (e.g., patient's name, primary ICD-10) and the remaining 181 are disease-specific data. On the other hand, from these 230 data fields only 23 are free-text entries (e.g., patient's name), the remaining 217 are structured data (e.g., flow-cyto-immunophenotype).

7.5.1 BASIC PRINCIPLES

Registers can be freely written, read, and modified, but can only be deleted if they are not associated with a schema.

Schemas are always created in a register-bound manner, each schema also having a version number one greater than the previous highest version number associated with

that register. The initial schema version number is 1. Schemas cannot be edited, only the logical value – indicating their active status – can be modified. Schemes have a validity period in addition to the version number and, due to transition periods, more than one scheme can be active at the same time. Schemas can be deleted if the schema to be deleted does not have a register entry. If a schema with the highest version number associated with a register is deleted, the version number is also reduced.

Each **schema** may contain **index** fields to speed up queries and statistics. Index fields are specified by pointing to an input field of a register schema e.g., 'Section 1 -> Chapter 2 -> Input field 3'. When an entry is added or modified, the system copies the value of the index fields pointed to into native database fields, on which indexes can be built.

Each **schema** may contain **validation** rules for each field, describing when a submitted register entry matches the schema.

A register **schema** may also contain **consolidation** fields that specify when two entries can be merged. If no consolidation fields are specified, all entries are added to the register as new entries, no merging is performed.

Register entries can be freely written, read, modified, and deleted, but they are subject to a variety of complex logic operations. The register entry indicates which register corresponds to which schema, and which schema is to be validated and consolidated.

The infrastructure of the ClinRegSys engine can be seen on the following figure:

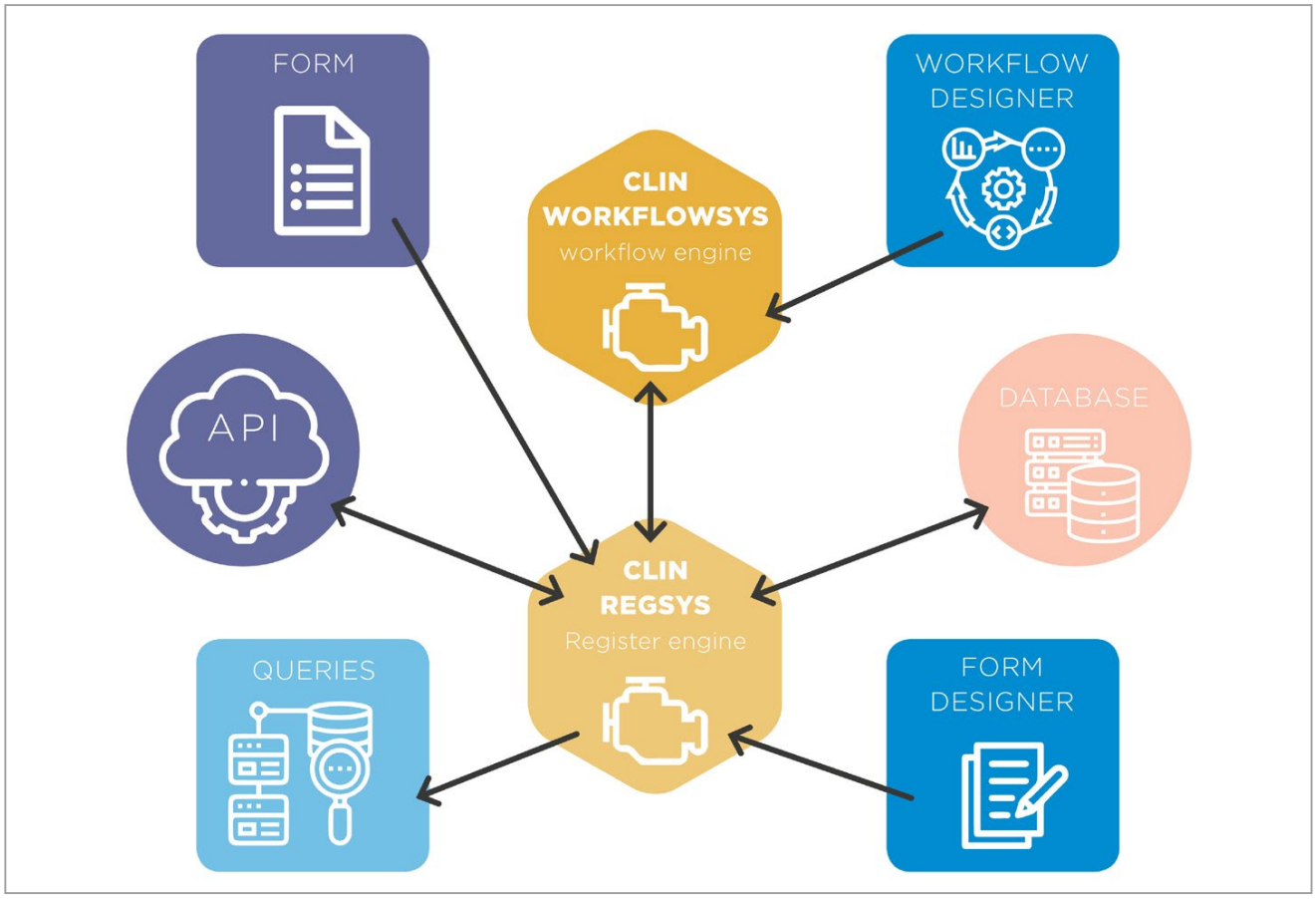


Figure 12: Engine infrastructure in a quality patient registry
Source: eHealth Software Solutions 2023

7.5.2 ADDING AN ENTRY TO A REGISTER

When a new register entry is added, a complex process is executed which performs several operations. To add an entry, the schema – and the register it specifies – must be active. It is also a condition that the content of the register entry satisfies the validation requirements specified in the schema, which are basically the following:

- validation of the content of fields against the rules defined in the schema at field level,
- checking the number of reusable sections.

In the figure one can see the detailed complex workflow of inserting a new entry into a register.

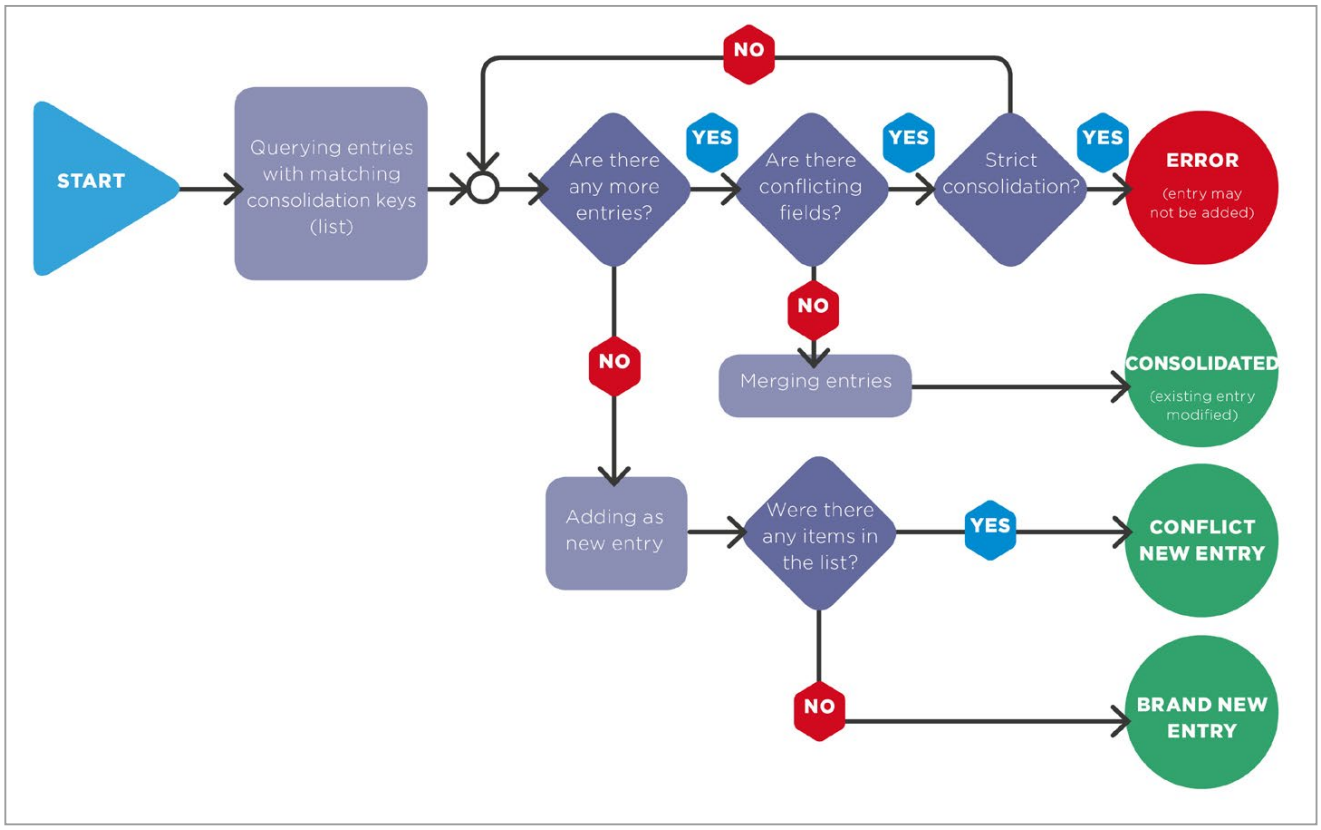


Figure 13: Workflow of inserting a new register entry to a register
Source: eHealth Software Solutions 2023

The validation is followed by the generation of the so-called consolidation key, which is used to assemble the entries containing the same key data. The key is generated based on the consolidation field list defined in the schema, first a HASH value is generated using the SHA256 function and then a key is generated from the HASH value using Base64 encryption.

If the consolidation key of an existing register entry and a new register entry to be inserted are identical, ClinRegSys checks whether the two records can be merged. The two records can be merged if there are no conflicting fields. If there are conflicting fields, two separate register entries are created, and if there are no conflicts, the two entries are combined into one consolidated entry.

Finally, the system updates the index values based on the new register entry according to the schema.

7.5.3 HISTORY OF ENTRIES

The ClinRegSys engine stores all register entry modifications in a versioned and structured way. Modifications can be of the following types:

- Creation
- Consolidation (concatenation)
- Modification (editing)
- Deletion

In the case of consolidation and modification, the changed fields are also stored with the old and new values.

7.5.4 QUERIES

The ClinRegSys engine includes a dynamic filtering subsystem to build and run queries according to specified conditions. The conditions are stored in register schemas and include name, description, query base and parameters. In all cases, the parameters have a title and description, and the type of value expected as a parameter. Compiled queries can be saved and loaded.

The system provides the possibility to export the data in the registers in tabular form. The export is performed by first selecting a group of entries using queries, followed by selecting the columns of the table. The result of the export is a table with columns as described above and values based on the filtering specified.

In the flowing figure the query designer and the result export can be shown:

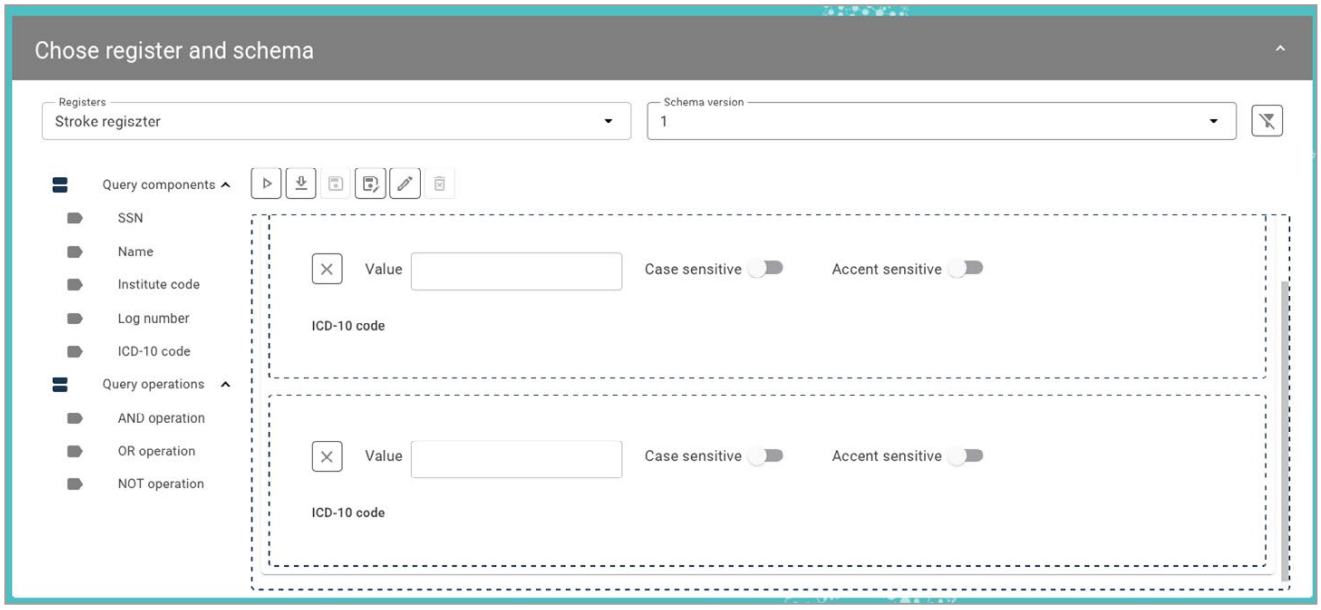


Figure 14: User interface of the query of a register
Source: eHealth Software Solutions 2023

7.5.5 DYNAMIC ENTRIES

In addition to the data entry and query user interface, the register engine shall provide an API interface for data upload and query. This API shall allow processes to run on receipt of a register entry based on its contents, which perform further actions depending on the received entry.

- They may request additional information from the submitting source.
- They may request clarification from the submitting source (considering the data already stored in the register).

Dynamic communication is handled by ClinWorkFlowSys, a health process engine that works in conjunction with the register engine.

In the National Registry of Haematology Diseases lots of haematology data are derived from laboratory result. The register store them as proprietary formatted JSON structure, which contains all the necessary field for a LOINC standard interface.

The registry workflow is semi-dynamic, which means it depends on the disease and previous data (history-dependent), but all scenarios have to be determined in advance.

7.5.6 LOCALISATION OF THE REGISTER ENGINE

The register forms can be localised, every single title, description or hint can be specified in multiple languages. Input fields of the register schemas can be general text, numeric, date/time input fields or dropdown lists. For the latter, localisation of the dropdown items is supported even for large datasets like health classification lists (e.g. ICD-10). By using dataset localisation, multiple language version descriptions of the entries can be specified using the same (common) code as a key. For example for the ICD-10 code the description is 'Cholera due to Vibrio cholerae 01, biovar cholerae' in English and 'Cholera durch Vibrio cholerae O:1, Biovar cholerae' in German. The form and all of its contents are displayed to the user in their selected language, if a list entry does not contain the requested language version of the description, it is not displayed.

7.5.7 DATA STRUCTURE OF REGISTER ENTRIES

The register entries are stored as JSON object. In each entry there are the identifier of the referred register schema entry. The figure shows an example of the multi-language input form as well as the stored structured of the entered data.

The registry entries stored as one object, and due the ClinRegSys engine and the database management system they are searchable, filterable, can be sorted. They are independent of the modification of content structure, and the structured or free text data are ready for ontology-based process.

The following figure shows the structure of a data-entry form descriptor and a JSON object of a register entry.

Figure 14: JSON data structure of a register entry form
Source: eHealth Software Solutions 2023

7.5.8 REFERENCES

In Hungary, the ClinRegSys engine is used by the National Haematological Diseases Registry (NHBR), the National Infectious Diseases Registry (NFBR), the National Stroke Registry (NSR) and the National Affective Diseases Registry (NABR). Two or two registries are operated by two national health institutions. Registries within an institution may also be linked to each other to support cross-registry research.

7.6 SWEDISH API-EFFORTS

The Swedish Agency for Digital Government (DIGG) regards API:s as a central issue if the public sector should be able to contribute to the development of a common administrative digital infrastructure that promotes innovation and broad societal benefit through better conditions for efficient data sharing and data use.

Using data from many different sources can be a time-consuming and expensive task. One of the drivers for the cost are poorly written/undocumented API:s as well as a wide variety of solutions for more or less the same functions. DIGG has therefore created an API Playbook that aims to create a kind of “soft” standard for API-management, i.e. how the API:s are measured regarding their maturity, maintained and developed over time.

The agency has also released an API-profile that gives technical advice on how to design an API.

Artefacts:
API Playbook - https://dev.dataportal.se/api-playbook
REST API-profile - https://dev.dataportal.se/rest-api-profil

8. VALUE MODELLING AND ESTIMATIONS

Accurate metadata annotation, which involves providing detailed and accurate information about data, can add significant value to datasets and facilitate effective data management, analysis, and utilisation.

Accurate metadata annotation can help users quickly identify and locate relevant datasets, understand their contents, and determine their suitability for specific purposes. This can save time and effort in searching for and selecting appropriate datasets, resulting in increased productivity and more efficient data utilisation.

Metadata annotation can provide critical information about the quality, completeness, and reliability of data, helping users assess the data's reliability and fitness for use. Accurate metadata annotation can also describe data structure, format, and schema, facilitating data integration, analysis, and interoperability. This can lead to improved data quality, usability, and trustworthiness, which are essential for making informed decisions based on data.

It can provide details about data ownership, licensing, and usage rights, which can facilitate data sharing and collaboration among different stakeholders. Accurate metadata annotation can also include documentation on data provenance, versioning, and update history, enabling users to understand the lineage and history of data, which is important for reproducibility and accountability in research and analysis.

Metadata annotation can include standard vocabularies, taxonomies, or ontologies that provide a common language for describing data, facilitating data interoperability and standardisation. This can enable data integration, exchange, and reuse across different systems, platforms, and domains, promoting consistency and harmonisation in data management and analysis.

Annotation can include information about data governance policies, procedures, and guidelines, helping users understand and comply with data governance requirements. Accurate metadata annotation can also include data privacy and security information, enabling users to assess and manage data risks, comply with relevant regulations, and protect sensitive data.

Accurate metadata annotation can potentially yield several financial benefits, as it can streamline data management processes by facilitating data discovery, retrieval, and integration, which can save time and effort. This can result in reduced operational costs associated with data acquisition, preparation, and processing. It can

- enable faster and more efficient data analysis, as users can quickly identify relevant datasets and assess their suitability for specific purposes. This can result in increased productivity, faster decision-making, and better resource allocation.
- improve data quality by providing information about data provenance, quality, and completeness. Higher quality data can lead to improved decision-making, reduced errors, and better outcomes, which can have financial benefits in terms of improved efficiency, customer satisfaction, and competitive advantage.

- facilitate data sharing and collaboration among different stakeholders, leading to i creased opportunities for joint research, innovation, and business partnerships. This can result in cost sharing, increased revenue through new collaborations, and access to new markets or customers.
- help organisations comply with data governance policies, regulations, and industry standards, which can prevent potential fines, legal liabilities, and reputational risks associated with non-compliance.
- enable organisations to better understand the value and potential uses of their data, leading to increased opportunities for data monetisation through data licensing, data products, or data-based services.

Accurate metadata annotation in healthcare can

- lead to improved patient outcomes, reduced medical errors, and more effective treatment plans. For example, accurate metadata annotation of electronic health r cords (EHRs) can help healthcare providers better identify and manage patients with chronic conditions, reduce unnecessary tests and procedures, and optimise treatment plans. This can result in cost savings through reduced hospital readmissions, improved patient satisfaction, and more efficient healthcare delivery.
- help reduce medical errors by improving data accuracy and integrity, leading to cost savings associated with reduced adverse events, rework, and legal expenses.
- facilitate better patient care by enabling more accurate diagnosis, treatment planning, and monitoring. This can result in improved patient outcomes, such as reduced hospital readmissions, better disease management, and improved patient satisfaction. Improved patient outcomes can lead to cost savings through reduced healthcare utilisation, improved patient retention, and increased reputation for healthcare providers.
- support population health management initiatives, such as disease surveillance, outbreak detection, and public health interventions. Timely and accurate metadata annotation of health data can enable early identification of disease trends, effective interventions, and targeted population health programs, leading to improved health outcomes and potential cost savings from reduced disease burden and healthcare costs.
- support precision medicine and personalised treatment plans. Accurate metadata annotation can enable more precise identification of patient cohorts for specific treatments or interventions, leading to optimised treatment plans, reduced trial-and-error approaches, and potentially cost-effective treatment strategies.
- enable better integration and analysis of diverse health data sources, facilitate data sharing and collaboration, and enhance research reproducibility. This can lead to faster research outcomes, improved research efficiency, and potential cost savings in research and development efforts.
- lead to more efficient data analysis, improved research outcomes, and faster time to market for new products or services. For example, accurate metadata annotation of scientific data can enable better data integration, analysis, and interpretation, leading to more robust research findings and accelerated innovation.

9. PROPOSED RESEARCH AND INNOVATION TOPICS

9.1 DOMAIN SPECIFIC DESCRIPTION OF DATA NETWORKS LINKED WITH SEMANTIC NETWORKS.

To explore methods for seamlessly integrating semantic networks with data networks, enabling domain-specific knowledge representation that enhances data integration, interpretation, and knowledge management within the chosen domain.

9.2 DOCUMENTATION AND RECORDING STANDARDS FOR CODED VS. FREE TEXT INPUT.

To comprehensively assess and compare the quality, efficiency, usability, interoperability, semantic analysis potential, and establish standardisation guidelines for coded (structured) and free text (unstructured) input methods across various domains, aiming to inform the choice of input method and promote data reliability and efficiency in diverse applications.

9.3 DOCUMENTATION AND RECORDING STANDARDS FOR THE TWO TYPES OF KNOWLEDGE

Decisions made by health professionals are influenced by the types of memories and knowledge, both explicit and tacit, which are predominantly based on experience. The objective is to explore possibilities for the introduction of symbols and images, sketches generated by image-generating AI solutions.

9.4 DATA VALUATION STUDY

Following the transformation of the health sector, strengthening the citizens' side of the development of the European Health Data Space with projects that link the health value benefits demonstrated to individuals and communities in a data-driven value chain that can be implemented within the EHDS framework and infrastructure to promote personalised health.

CONCLUDING STATEMENT

The data age we have entered with the new millennium will bring us decades of health, according to researchers in many disciplines. Our own data copy and the data footprint of our environment play a crucial role in this. We can harness this emerging virtual data space to explore fundamentally new ways of understanding life processes by exploring it with the emerging tools of mathematics. Various artificial intelligence and machine learning tools are bringing us forms of learning that reveal knowledge about ourselves and our humanity that has hitherto been barely accessible to us. This brings within our reach an increasing emphasis on health rather than disease management, with health navigation systems, such as traffic navigation systems, available through personal digital devices to support our everyday healthy living.

In this White Paper, we call attention to the fact that for this to happen, data must be empowered to be our primary resource for creating health every day. This requires the development of methodologies and processes that, like trajectories in cartography, map the complex, interconnected and interrelated living systems of people and their environments in a well-defined way in a multidimensional virtual data space.

In addition, in order to effectively support and guide the human life process, life situation-based knowledge graphs and their associated semantic networks must be used to create an information model that can form the basis for the unique pattern recognition and prediction capabilities of artificial intelligence.

In doing so, we will also provide a systems approach to health and medicine in which we can apply the tools of mathematics and physics to cognition, modelling and the development of supporting tools.

We wish you successful navigation and exploration in the redefined health data space!

REFERENCES

1. McKinsey & Company, 2017 (https://www.mckinsey.com/industries/life-sciences/our-insights/real-world-evidence-from-activity-to-impact-in-healthcare-decision-making#/) based on Health Policy Brief: The Relative Contribution of Multiple Determinants to Health Outcomes, Health Affairs, August 21, 2014.
2. FEAM Forum Annual Lecture 2022: Digital Health and AI on 26 October, https://www.youtube.com/watch?v=c3dw4ImoUNc
3. Zeinab M. Mamdouh, Elisa Anastasi and Ahmed A. Hassan et al.: Why the way we define diseases prevents innovation and precision medicine. DrugRxiv. 2023. DOI: 10.14293/S2199-1006.1.SOR-.PPCFYDY.v1
4. Bai et al: STG2Seq: Spatial-temporal Graph to Sequence Model for Multi-step Passenger Demand Forecasting, 2019. https://doi.org/10.48550/arXiv.1905.10069
5. Lantos 2022, based on Ackoff 1989 in: The Next Era in Global Health by Copenhagen Institute for Futures Studies, 2020.
6. GRADE = Grading of Recommendations Assessment, Development, and Evaluation (https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade/)
7. The Next Era in Global Health by Copenhagen Institute for Future Studies, 2020.
8. Witte EH, Stanciu A, Boehnke K: A New Empirical Approach to Intercultural Comparisons of Value Preferences Based on Schwartz's Theory. Front. Psychol. 11:1723. 2020
9. Prainsack, B et al.: Data solidarity: a blueprint for governing health futures. The Lancet Digital Health, Volume 4, Issue 11, e773-e774. 2022
10. The Next Era in Global Health by Copenhagen Institute for Futures Studies, 2020.
11. Verheij, R.A., Curcin, V., Delaney, B.C., McGilchrist, M.M. Possible sources of bias in primary care electronic health record data use and reuse. Journal of Medical Internet Research: 2018, 20(5), e185, https://pubmed.ncbi.nlm.nih.gov/29844010/
12. https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/3-interoperability-layers
13. InteropEHRate project. Simone Bocca, Gábor Bella, Yamini Chandrashekar, 2022.
14. Glass, Michael & Rossiello, Gaetano & Gliozzo, Alfio. (2021). Zero-shot Slot Filling with DPR and RAG.
15. A. Meloni, S. Angioni, A. Salatino, F. Osborne, D. Reforgiato Recupero and E. Motta, "Integrating Conversational Agents and Knowledge Graphs Within the Scholarly Domain," in IEEE Access, vol. 11, pp. 22468-22489, 2023, doi: 10.1109/ACCESS.2023.3253388.
16. Demetriadis, S., Dimitriadis, Y. (2023). Conversational Agents and Language Models that Learn from Human Dialogues to Support Design Thinking. In: Frasson, C., Mylonas, P., Troussas, C. (eds) Augmented Intelligence and Intelligent Tutoring Systems. ITS 2023. Lecture Notes in Computer Science, vol 13891. Springer, Cham. https://doi.org/10.1007/978-3-031-32883-1_60

17. The Personal Health Knowledge Graph Workshop.4 May 2021, Virtual.
18. https://phkg.github.io/
19. https://vimeo.com/529328116/732160e4b4
20. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. J Med Internet Res. 2018 May 29;20(5):e185. doi: 10.2196/jmir.9134
21. Kuchinke W, Ohmann C, Verheij RA, van Veen EB, Arvanitis TN, Taweel A, Delaney BC. A standardised graphic method for describing data privacy frameworks in primary care research using a flexible zone model. Int J Med Inform. 2014 Dec;83(12):941-57. doi: 10.1016/j.ijmedinf.2014.08.009. Epub 2014 Sep 3.
22. https://www.aihw.gov.au/about-our-data/accessing-data-through-the-aihw/meta-data-standards
23. https://meteor.aihw.gov.au/content/268284
24. https://martinfowler.com/articles/data-mesh-principles.html
25. https://www.mydata.org
26. https://www.datavaults.eu